Single-cell Morphological Data Reveals Signaling Network Architecture

by

Oaz Nir

B.S., Mathematics, B.A., English
Duke University, 2005

SUBMITTED TO THE HARVARD-MIT DIVISION OF HEALTH SCIENCES AND TECHNOLOGY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY IN THE FIELD OF MATHEMATICS AND HEALTH SCIENCES AND TECHNOLOGY
AT THE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

FEBRUARY 2010

Signature of Author:_____

Harvard-MIT Division of Health Sciences and Technology
MIT Department of Mathematics
December 14, 2009

Certified by:_____

Bonnie Berger, PhD
Professor of Applied Mathematics
Thesis Supervisor

Accepted by:_____

Ram Sasisekharan, PhD
Director, Harvard-MIT Division of Health Sciences and Technology
Edward Hood Taplin Professor of Health Sciences & Technology and Biological Engineering

Accepted by:_____

Michel X. Goemans, PhD
Chairman, Applied Mathematics Committee

Accepted by:_____

Bjorn Poonen, PhD
Chairman, Mathematics Department Committee on Graduate Students

Table of Contents

Acknowledgment

The author thanks his thesis advisor, Professor Bonnie Berger. He also thanks the other members of his thesis committee, Professors Norbert Perrimon and Daniel Kleitman. This work would not have been possible without the biological contributions of Dr. Chris Bakal. The author wishes to thank the following individuals for helpful conversations over the course of his work on this thesis: Uri Laserson, Michael Baym, Rohit Singh, Nathan Palmer, and Patrick Schmid (all of whom are graduate students at the time of this writing). He wishes to thank Professors Alexander van Oudenaarden and Tommi Jaakkola, who taught the two MIT courses which proved most useful for this thesis work. Going further back, he thanks Professor John Harer of Duke University, who first introduced him to the field of computational biology.

The author provides most gracious thanks those that provided financial support to this work: the MIT Presidential Fellowship program, namely to Akamai, the Harvard MIT Division of Health Sciences and Technology through the MEMP Fellowship, the NIH through the HST BIG Training Grant in Bioinformatics. and especially to the Department of Energy for its Computational Science Graduate Fellowship.

This publication was made possible by Grant Number T32 HG002295 from the National Human Genome Research Institute (NHGRI). Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the National Human Genome Research Institute (NHGRI). This work was supported by the Department of Energy Computational Science Graduate Fellowship Program of the Office of Science and National Nuclear Security Administration in the Department of Energy under contract DE-FG02-97ER25308.

Single-cell Morphological Data Reveals Signaling Network Architecture

by

Oaz Nir

Submitted to the Harvard-MIT Division of Health Sciences and Technology and MIT Department of Mathematics on December 14, 2009 in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy at the Massachusetts Institute of Technology.

Abstract

Metastasis, the migration of cancer cells from the primary site of tumorigenesis and the subsequent invasion of secondary tissues, causes the vast majority of cancer deaths. To spread, metastatic cells dramatically rearrange their shape in complex, dynamic fashions. Genes encoding signaling proteins that regulate cell shape in normal cells are often mutated in cancer, especially in highly metastatic disease. To study these key signaling proteins in locomotion and metastasis, we develop and validate statistical methods to extract information from high-throughput morphological data from genetic screens.

Our contributions fall into three major categories. 1) To define and apply robust statistical measures to identify genes regulating morphological variability. We develop and thoroughly test methods for measuring morphological variability of single-cells populations, and apply these metrics to genetic screens in yeast and fly. We further apply these techniques to subsets of genes involved in cellular processes to study genetic contributions to variability in these processes. We propose new roles for genes as suppressors or enhancers of morphological noise. We validate our findings on the basis of known gene function and network architecture. 2) To perform inference of protein signaling relationships by utilizing high-throughput morphological data. We apply machine-learning techniques to systematically identify genetic interactions between proteins on the basis of image-based data from double-knockout screens. Next, we focus on RhoGTPases and RhoGTPase Activating Proteins (RhoGAPs) in *Drosophila.*, where by using basic knowledge of network architecture we apply our techniques to detect signaling relationships. 3) To integrate expression data with high-throughput morphological data to study the mechanisms for determination of cell morphology. We utilize morphological and microarray data from fly screens. By comparing expression data between control treatment conditions and treatment conditions displaying morphological phenotypes (e.g. high population variability), we identify genes and pathways correlated with this class distinction, thereby validating our previous studies and providing further insight into the determination of morphology.

A key challenge in systems biology is to analyze emerging high-throughput image-based data to understand how cellular phenotypes are genetically encoded. Our work makes significant contributions to the literature on high-throughput morphological study and describes a path for future investigation.

Thesis Supervisor: Bonnie Berger
Title: Professor of Applied Mathematics

Chapter 1:
Introduction

Metastasis, the migration of cancer cells away from the primary site of tumorigenesis and the subsequent invasion of secondary tissues, is the cause of the vast majority of patient deaths due to cancer. In order for metastatic cells to spread throughout the body they must dramatically rearrange their cell shape and morphology in complex and dynamic fashions. As such, genes which encode for signaling proteins that regulate cell shape in normal cells are often targets of multiple mutations in cancer, and especially in highly metastatic forms of the disease. In order to study these key signaling proteins in locomotion and metastasis, we develop and validate statistical methods to extract information from high-throughput morphological data from genetic screens. More specifically, we develop techniques to identify genetic components of cellular morphological variability, utilize morphological data to identify genetic interactions and perform signaling pathway inference, and integrate morphological and transcriptional data to study determination of cell morphology.

Background and Significance

Understanding metastasis through systems-level analysis of signaling networks

Cancer is soon to become the leading cause of death in the United States as mortality rates for the disease have remain largely unchanged for the past 50 years, while death-rates for cardiac, cerebrovascular, and infectious diseases have markedly declined [1]. As nutrients and space become limiting at the primary site of tumorigenesis, a small population of cancer cells become metastatic, meaning these cells acquire the ability to migrate from the tumor mass and invade other tissues. The conversion of a locally-growing tumor to one that gains the ability to metastasize is a critical and damaging point in cancer progression, as 90% of cancer patient deaths are due to the effects of tumor cells that have founded colonies in tissues distant from the primary site [2]. Metastasis is a biologically complex process that still remains poorly understood, but is known to involve dramatic changes in the morphology of tumor cells.

In all eukaryotic cells, Rho family guanosine triphosphatases (GTPases) such as Rho, Rac, and Cdc42, are the master regulators of cell morphology as these proteins dynamically integrate a vast spectrum of upstream signals and directly control the actin and microtubule cytoskeletons, cell-cell or cell-matrix adhesion, vesicular trafficking, and cell polarity [3]. For example, cell migration is due to the coordinated actions of Rho and Rac GTPases on actin organization. In order to generate the driving force required for motility, Rac-type GTPases at the leading edge of cells promote actin polymerization and the formation of protrusive lamellipodia, while at the trailing edge of cells Rho-type GTPases control the actomyosin machinery in order to stimulate contraction [4]. During migration Rac and Rho also act to regulate the formation of integrin based adhesions at the leading edge, while coupling cell contraction at the rear of the cell to adhesion disassembly [4]. Defects in any of these processes lead to an inability of cells to migrate in an efficient manner. Dysregulated Rho signaling has been widely implicated in the metastasis of many tumor types, especially breast tumors [5].

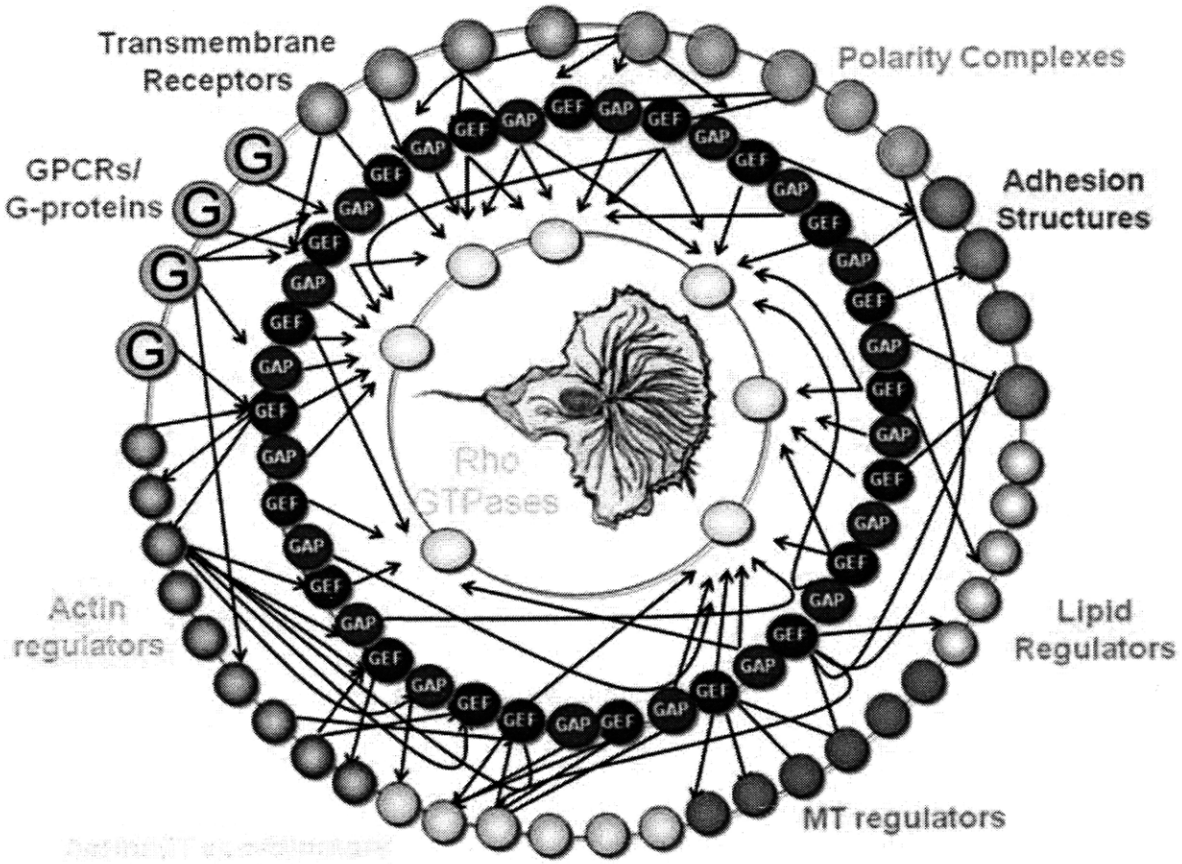


Fig. 1. Schematic of signaling networks governed by GTPases and involved in cell locomotion and morphogenesis. As network hubs, GTPases activity is regulated by upstream pathways (e.g. GPCRs) and, in turn, regulates downstream (e.g. MT regulators) activity.

Spatio-temporal regulation of key signaling proteins is responsible for the morphology of a single cell. Many signaling proteins must act both upstream and downstream of specific Rho GTPases in spatially distinct subcellular local networks to translate extracellular signals to changes in GTPase activation and ultimately in cellular morphology [6, 7]. Twenty genes encoding different members of the Rho family have been identified in the human genome, and it is assumed that each one acts as a molecular switch to control distinct biochemical pathways [8]. Like all

regulatory GTPases, these proteins exist in an inactive GDP-bound conformation and an active GTP-bound conformation. The GDP/GTP cycling of GTPases is tightly controlled by a large family (85 in mammals) of GEFs (guanine nucleotide-exchange factors) that increase GDP/GTP exchange rates. Regulation of GEFs is not well understood, and it is crucial to develop further methods to understand the spatio-temporal activation of GTPases. An equally large family of GAPs (GTPase-activating proteins) that activiate the intrinsic GTPase activity of Rho GTPases has been identified [9]. Although they are likely to down-regulate GTPase signalling, even less is known about how they are recruited and activated than in the case of GEFs.

In an attempt to define the biochemical pathways activated by Rho GTPases, many groups have used yeast two-hybrid selection and affinity chromatography techniques to identify cellular targets of Rho, Rac and Cdc42 [10-14]. More than 50 potential targets have been identified to date, and a current major task is to define their individual roles [7]. The *Drosophila* BG-2 cell line is a particularly useful model system to study GTPase signaling pathways and to study local networks to control cell morphology. BG-2 cells display a high degree of cellular motility and exhibit many of the morphological characteristics of mammalian fibroblasts and epithelial cells, including the formation of integrin-based adhesions, polarized lamellipodia, and coordinated retraction of the cell body [15, 16].

In sum, identifying signaling relationships of GTPases is instrumental toward further understanding of regulation of cell morphology and locomotion. Classical biochemical techniques for validating *in vivo* interactions are time-consuming and expensive, while more traditional high-throughput techniques, such as yeast two-hybrid experiments, provide poor predictions. There is a profound need to acquire high-throughput morphological data in a fast

and cheap manner, and to develop appropriate methods to extract information from this data to study the key signaling relationships regulating cell locomotion and morphological change.

High-throughput single-cell image acquisition and quantification of morphology

Image-based automated technologies and acquisition of high-throughput quantitative imaging data is a recent development and these technologies have been applied to quantify shape, DNA morphology, and the subcellular localization of organelles or proteins [17-19]. Additional groups have performed high-throughput image acquisition in cell culture across different species and cell lines. Typically these experimental tools were used in the context of chemical or genetic screens, generally aided by RNAi where single cells from hundreds-thousands of different treatment conditions have been analyzed [20-28].

In these studies high-throughput image acquisition involves a sequence of steps. First, raw images are acquired using automated microscopy; next, segmentation is performed to identify single cell images; finally, quantification of single-cell morphology is performed. Typically, dozens-hundreds of morphological features are defined and measured for each single cell. Thus a morphological signature is obtained for each single cell, and by extension a quantitative morphological description is obtained for each treatment class in the chemical or genetic screen.

Here we describe in some detail the methods of a previous study of Bakal et al. [19] in *Drosophila* BG-2 cells, which forms an important data source for our computational analysis. Instead of completely automating segmentation, researchers developed a software application (CellSegmenter) for computer-assisted segmentation. Over the course of ~10 months, 12,601 individual cell segments falling into 273 treatment conditions (also called TCs) defined by

overexpression of knockout of a single gene as well as control TCs were generated using CellSegmenter. Automated image analysis algorithms were developed to compute 145 mathematical values (features) for each of these segments from the cell segment image created by CellSegmenter and the original GFP intensity image. The features were designed to interrogate aspects of the overall geometry and size of the cell segments, the stochastic GFP label intensity, and the statistical distribution and 'texture' of this intensity with relation to cell geometry. Other features measured attributes of the shape of the cell boundary as rendered by the cell segment, including the number, size, shape, and distribution of processes and undulations of the boundary as analyzed at both a small and a large scale. The feature set also included a number of previously published features reported to be useful for analyzing the cytoskeletal behavior of cells.

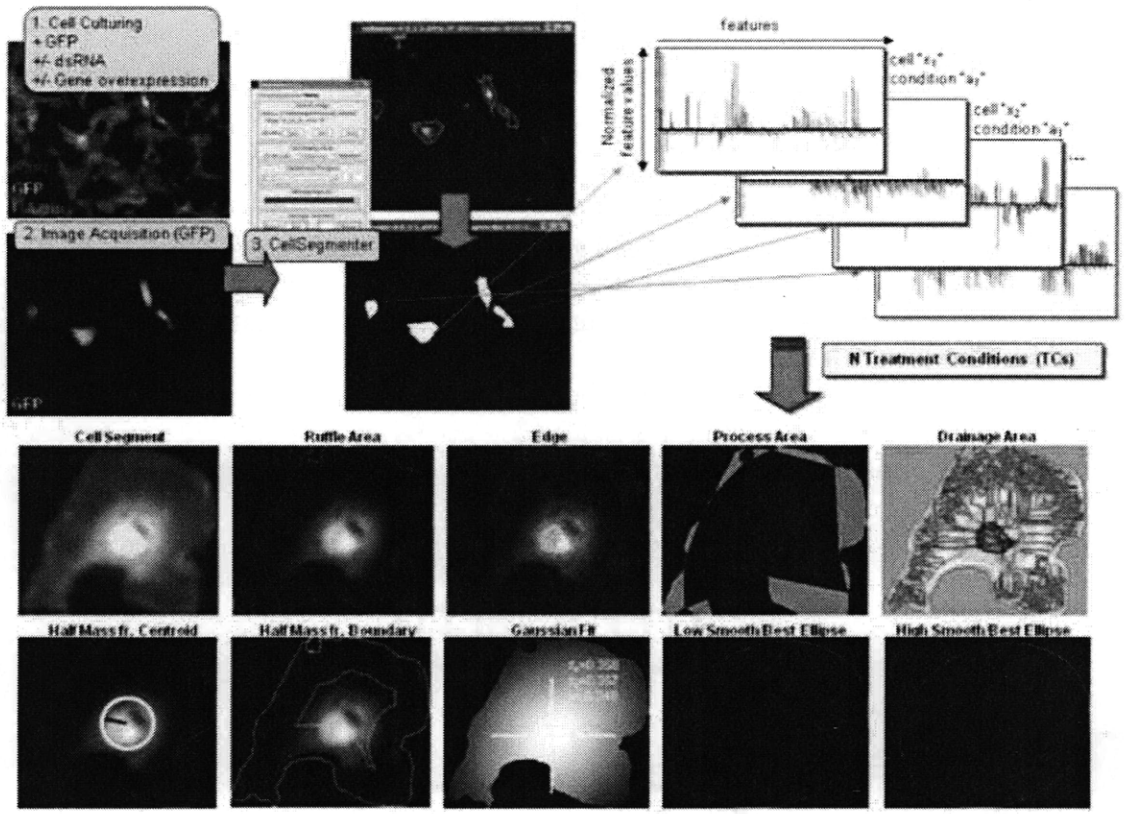


Fig. 2. Acquiring morphological signatures from complex images. Cultured *Drosophila* BG-2 cells were transfected with plasmids encoding GFP and either cotransfected with plasmids encoding red fluorescent protein–tagged proteins or incubated in the presence of dsRNA for 4 days (Top, left). Images of GFP-labeled cells were acquired by standard fluorescence microscopy, and individual cell images with clear and complete boundaries were selected (Top, middle). 145 different features relevant to cell morphology and GFP signal intensity were derived from individual cells (Top, right). Examples of features computed from individual cell images (Bottom panel). Figure adapted from [19].

While the feature analysis generated 145 numerical features corresponding to aspects of morphology for each cell segment and thus provided immense information about cell morphology, these features had complex relations to each other and unclear biological interpretation. It is, in general, therefore necessary to perform dimensionality reduction on raw

morphological data. In this case, Bakal et al chose to use neural networks to train classifiers for certain archetypal cell morphologies, and then to use normalized scores of these classifiers as the basis for reduced-dimensional space.

Finally, mean scores for each TC in this reduced space were used to cluster all TCs. Hierarchical average linkage clustering was performed using uncentered Pearson Correlation Coefficients as the distance measure. Enrichment statistics for the final clustering were computed against functional category information derived from Gene Ontology [29]. The clustering analysis yielded 41 total clusters, which in some cases correspond to known morphological processes.

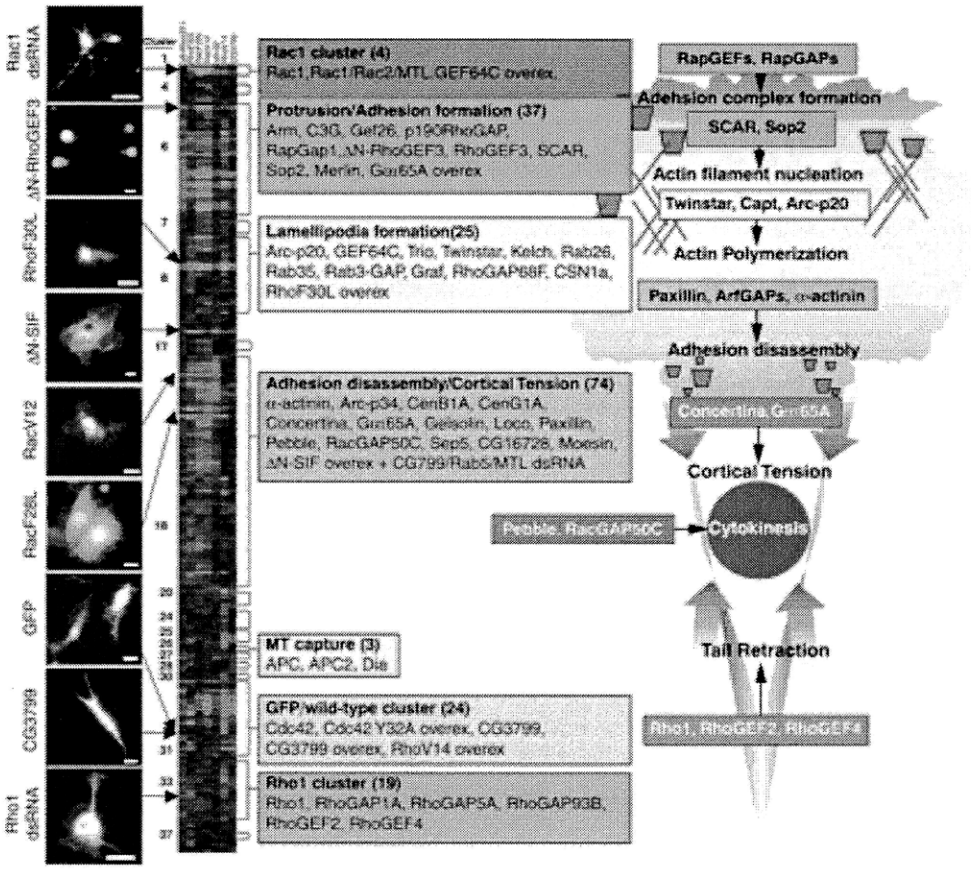


Fig.3. Identification of local networks morphological data. TCs were subjected to hierarchical clustering; all 41 multigene clusters are identified in brackets on the right-hand side of the clustergram. For some clusters, prominent TC are listed, and the number of TCs within these clusters is indicated in parentheses. Examples of individual cells and their positions in the clustergram are shown on the left-hand side of the clustergram. Based on their gene membership, 7 of the clusters were determined to have specialized roles in cell morphology, as shown in the 7 boxes. Figure adapted from [19].

Measurement of morphological variability

Techniques for measuring morphological variability of a population of single cells, belonging to a single treatment condition in a genetic screen, are at present limited in scope. Data analysis of the high-content datasets obtained from automated image acquisition typically begins by reducing the dimensionality of the space of morphological features [30]. Subsequently, data analysis has commonly been performed by averaging the results from single-cell measurements for each treatment condition to derive mean scores for each condition or by performing clustering analysis [19, 31-33]. But mean scores or clustering fails to fully capture the wealth of single-cell morphological data these studies provide. Recently, Levy et al. [34] quantified population variability by studying the variance of composite features in a genetic screen in yeast, and showed that knockout of genes with high network connectivity (i.e. network hubs) tends to increase observed morphological noise. However, Levy et al. do not analyze properties of genes which, when knocked out, decrease morphological noise, nor do they measure genetic contributions to morphological noise in specific cellular processes or study morphological variability in higher eukaryotes. Slack et al. [35] quantified population variability in a chemical screen in HeLa cells by viewing populations as mixtures of phenotypically distinct

subpopulations and viewing chemical response as a redistribution of the relative subpopulation proportions. This approach was successful in classifying drugs according to mechanism of action. Consideration of genetic modulation of morphological noise in specific cellular processes, however, has been limited to apoptosis [36]. Quantifying noise in a non-lethal cellular process, where cell shape may change along multiple dimensions, is a much more complex problem than measuring presence/absence of cell death.

An additional challenge is to formulate a general model for morphological variability that accounts not only for genetic variants (e.g. RNAi of a gene of interest) but for additional modulators of morphological variability, such as external condition (cell concentration, temperature, etc), cell type, and cell cycle variation. Models for variability/noise in transcriptional data have been developed [37-40], but to our knowledge no comprehensive model exists for morphological variability.

Inference of signaling pathways – from traditional data sources to morphological data

Techniques for inference of signaling relationships between proteins of interest on the basis of transcriptional and phosphoproteomic data from gene knockout experiments, particularly Bayesian networks, have been well developed over the past decade [41-49]. Microarrays are capable of measuring the expression level of thousands of genes simultaneously. Early analysis of microarray data focused on identifying clusters of genes that exhibit tightly coupled transcriptional response with respect to phenotypic classification or environmental stimulation. Soon thereafter, interest arose in discovering transcriptional "fingerprints" associated with phenotype. Because it is widely believed that oncogenesis and metastasis are primarily mediated

through transcriptional programs, much of that attention was focused on the analysis of tumor cells [50, 51].

A major challenge in systems biology over the last several years has been to move beyond these clustering and classification-oriented methodologies and develop a finer-grained, dynamic picture of how transcriptional response fits into the larger picture of cell signaling. As such, several studies over the past seven years have attempted to interpret the large volume of publicly available [52] transcriptional data to model signaling networks. Friedman et al. [41] described a method for learning Bayesian Networks from microarray data. These graphical models encode a full joint multivariate probability distribution and dependence structure over the interacting elements in a domain (e.g., observed transcription levels). The edges that connect various elements may indicate some of the causal influence structure in the domain. For example, an edge from a transcription factor to a target gene can encode the tendency for the particular transcription factor to enhance or repress transcription of the target. Properly calibrated, these models should accurately capture the interactions among a system of biological molecules. An important caveat to these studies is that levels of gene expression are assumed to correlate directly with levels of protein activation (i.e. high gene expression equals high protein activity), however given the multiple layers of regulation that exist to control gene/protein levels after transcription has occurred, this assumption can in some cases lead to misleading conclusions.

Similar techniques have also been applied to single-cell flow cytometry data in order to infer signaling relationships [53]. These methods provide a robust means for extracting systems-level information from raw biological data [54]. A severe limitation to cytometry analysis is that specific antibodies must be raised against each protein of interest in its activated and/or

nonactivated form, which is a time and labor consuming process. Such studies also require prior knowledge as to the components that make up the network.

Although it is challenging to extract signaling information from morphological data, as it provides a relatively indirect read-out of protein activity, unsupervised machine learning methods have been used to successfully identify groups of genes which co-regulate certain known morphological processes. We therefore hypothesize that high-throughput morphological data can be used to improve inference of signaling pathways based solely on microarray data. We focus our attention on treatment conditions defined by perturbations of GTPases and GTPase Activating Proteins (GAPs). GAP knockout yields cells with similar expression profiles as corresponding GTPase knockout. By comparing the quantitative morphological signatures of GAP RNAi treatment conditions to GTPase overexpression treatment conditions, we determine whether an analogous signal for morphological profiles may be used to identify signaling relationships.

Integrating transcriptional and morphological data

The literature on the analysis of transcriptional data is well-developed. Most relevant for our work here are methods to detect differential expression between two unpaired groups of treatment conditions [55-57]. A complication in microarray data analysis for the research community has been the proliferation of alternative methods for data normalization, values for cutoff parameters, and, even more basically, methods for determining significant differential expression. Accordingly, we use two alternatives for determining differential expression, one of which, t tests, is essentially the "industry standard, the other of which is Significance Analysis

for Microarrays (SAM) [58]. The SAM procedure is to compute a normalized coefficient of linear regression for each gene, relative to the class distinction, to determine FDRs for each value using resampling, and finally to identify up- and down-regulated genes by defining an FDR threshold.

A technique that is related to the analysis of differential expression, but is distinct from it, is gene set enrichment analysis (GSEA) [59, 60]. Using GSEA, one is able to detect overall enrichment for a gene set among either up- or down-regulated genes, even if the individual differential expression of the component genes is not statistically significant. The GSEA algorithm proceeds by keeping a running total of a statistic by traversing the list of gene probes, as ordered by correlation with the class distinction (e.g. beginning with the genes most up-regulated in the High Variability group as compared to control). The ES for the gene set is defined to be the maximum statistic encountered in this manner. Subsequently, a normalized ES (NES) is computed to account for differences in gene sets size and an FDR is computed on the basis of the each NES to account for multiple hypothesis testing.

With the emergence of high-throughput morphological screens, a key issue in systems biology is to integrate this data source with high-throughput transcriptional data from microarrays. In this work, we apply techniques from microarray data analysis to determine differential expression and gene set enrichment between group pairs defined by morphology-based class distinctions. We focus on morphological phenotypes corresponding to clustering of mean morphology, on the one hand, and quantification of population-level variability, on the other. We seek to study the mechanisms behind these phenomena by integrating analysis of morphological data with expression data. The use of morphology to define class distinctions for further study by transcriptional data has a long history, particularly in the study of cancer [55], but has not been

applied, to our knowledge, to high-throughput morphological data from a large-scale genetic screen.

Summary

High-throughput image-based morphological data can be acquired in a robust, cheap, and fast manner. The application of high-throughput automated image acquisition techniques has the potential to improve the understanding of cell signaling pathways involved in disease processes. While some preliminary work by other researchers has studied the role of network hubs in buffering variability in yeast, no methods have been developed to study morphological variability in specific cellular processes; nor has a genetic contribution to increasing morphological noise been studied; nor has any study of morphological variability at the cellular level in higher eukaryotes, where cellular phenotypes are significantly more complex than in yeast, been performed.. While approaches to predicting signaling pathways on the basis of transcriptional and phosphoproteomic data have been developed, methods to utilize morphological data for pathway inference are sorely lacking, as as methods for systemically identifying genetic interactions on the basis of high-throughput morphological data. Utilizing morphological data for this purpose is significantly more challenging than using these other data sources; yet, morphological data provides independent information about signaling relationships, thus motivating the development of techniques for its use. Finally, the integration of transcriptional and high-throughput morphological data from a genetic screen has not been performed to study the mechanistic basis for cell shape determination or to study the contribution of gene expression to key phenotypic distinctions.

Overview of Chapters 2-5

Here we provide a brief description of the work contained in the next three chapters of this thesis. For each chapter, we provide a brief rationale for the study as well as an overview of methods and results.

Chapter 2: To define and apply robust statistical measures to identify genes regulating morphological variability.

Rationale

Techniques for measuring morphological variability of a population of single cells, belonging to a single treatment condition in a genetic screen, are at present limited in scope. Data analysis of the high-content datasets obtained from automated image acquisition typically begins by reducing the dimensionality of the space of morphological features. Because data analysis routinely begins by averaging the results from single-cell measurements to derive mean scores for each condition or by performing clustering analysis, the wealth of single-cell morphological data is lost.

Overview of methods

Here we introduce methods for measuring genetic contributions to morphological variability for specific cellular processes. We develop a robust method for measuring population variability

more generally, and apply this method to genetic screens in both yeast and fly. The basis for the metric is relatively simple: a multi-dimensional analog of one-dimensional variance, applied to data that has been normalized by taking z-scores in each raw dimension and then reduced in dimensionality by using PCA. The benefit of applying a relatively simple methodology is increased confidence in the interpretation of results.

We apply our variability scoring procedure to study genetic contributions to morphological variability in specific cellular processes (protrusion/adhesion formation and adhesion disassembly/cortical tension in fly; septin ring formation in yeast). We validate our results based on known gene functions and network architectures for the processes under consideration. We find that the effects of genetic perturbations on morphological variability are explicable in many situations by the network architecture of the cellular process under consideration.

Overview of results

In the course of our analysis, we show that population-level morphological variability reflects the architecture of regulatory networks. Our methods and results extend the finding of Levy et al. [34] that knockout of network hubs tends to increase morphological variability. Here, we consider more intricate network architectures associated with regulation of complex cellular processes. Indeed, work in measuring single-gene transcription shows that perturbation of expression of genes with upstream products causes increased noise in the expression of downstream targets [40, 61]. We demonstrate repeatedly that perturbation of genes acting upstream in signaling pathways tends to increase morphological noise in the process mediated by the pathway to a greater extent than perturbation of genes acting further downstream in the

pathway. For example, in the case of septin ring assembly, knockout of HSL1 or HSL7 increases morphological variability to a far greater extent than knockout of SWE1, and perturbation of the upstream activators CLA4 and ELM1 results in increased variability.

Chapter 3: To perform inference of protein signaling relationships by utilizing high-throughput morphological data

Rationale

We hypothesize that high-throughput morphological data can be used to improve inference of signaling pathways based solely on microarray data. We focus our attention on treatment conditions defined by perturbations of RhoGTPases and RhoGAPs. In using expression data to perform inference, the core idea is that GAP knockout yields cells with similar expression profiles as corresponding GTPase knockout. By comparing the quantitative morphological signatures of GAP RNAi treatment conditions to GTPase overexpression treatment conditions, we determine whether an analogous signal for morphological profiles may be used to identify signaling relationships. More specifically, we first developed a systematic framework for identifying genetic interactions on the basis of high-throughput (single- and double-knockout) morphological data from an RNAi screen. We then applied this framework to infer RhoGAP/GTPase regulatory relationships by using prior knowledge of the basic structure of RhoGAP/GTPase signaling.

Overview of methods

Here we first acquire single-cell morphological data for TCs in the *Drosophila* BG-2 cell line defined by double-knockout of RhoGAPs. In particular, we acquire data for 90 additional TCs (all single and double-knockouts for 13 GAPs, excluding one case). For each TC, we acquire images for multiple single cells for a total of 6480 cells (an average of 72 for each TC). The same techniques used by Bakal et al. are applied to perform the double knockouts, culture cells, acquire cellular images, and extract geometric feature information.

We define a classification model for assigning a set of putative upstream TCs to a set of putative downstream TCs. This model is used to classify GAP knockout TCs onto the set of GTPase overexpression TCs. This analysis is repeated for double-knockout experiments. The classification model allows us to assign any new point, or set of points, in morphological space to one of several classes. In our case, we build two separate classification models. First, we use GTPase overexpression experiments as the downstream classes, and for each GAP knockout, we use the model to classify that knockout as belonging to one of downstream classes. Second, we again use GTPase overexpression experiments as the downstream classes, but this time use GAP double-knockouts as the set of upstream classes. This allows us to associate to each GAP a GTPase whose activity it is most likely to regulate. These predictions were compared to biologically validated interactions and non-interactions between GAPs and GTPases.

We develop a similar classification model using single knockouts as the set of downstream targets and double-knockouts as the set of upstream targets. Here, the terms "upstream" and "downstream" are not used literally but rather as descriptive terms for the model. By systematically identifying double-knockouts TCs that are morphologically similar to single-knockout TCs, we are able to construct putative hierarchies of action for GAPs. This is a way to quantitatively study the concept of genetic interactions, as well as the concepts of "party hubs"

and "date hubs" [62]. Indeed, the network of GAP/GEF/GTPase interactions is vastly interconnected, meaning that multiple GAPs and GEFs regulate the same GTPase. For example, suppose that RNAi of GAP1A and GAP5A most resembles RNAi of GAP1A as compared to RNAi of GAP5A alone, or in a stronger situation, any other GAP knockout. In this scenario, we say that the interaction of GAP1A is dominant over GAP5A, because once GAP1A is knocked-down, the further knock-down of GAP5A has no additional effect on morphology. Mechanistic explanations for such phenomena are provided by the party/date conceptualizations of network hubs; one possible explanation for the example above would that GTPase interaction with GAP1A is a necessary prerequisite for GTPase binding to GAP5A. By performing the clustering described here in a systematic way, we are able to organize the interaction hierarchy of GAPs.

Overview of results

The contributions of this section of the thesis are fourfold. The first contribution is to show the fact that high-throughput morphological data can be used in a systematic fashion identify genetic interactions. Second, we show the fundamental fact that with additional prior knowledge for the network structure, our framework can be used to identify signaling interactions successfully. Third, the computational framework presented here represents an initial approach to the problem that will serve as a basis for future enhancements. Fourth, and perhaps most intriguing, we showed that our classification model performs much better with both single- and double-knockout data versus only single-knockout data.

Chapter 4: To integrate expression data withhigh-throughput morphological data to study the mechanisms for determination of cell morphology.

Rationale

Here, we utilize the morphological data from the *Drosophila* genetic screen as well as microarray data from a similar screen. By comparing expression data between control treatment conditions and treatment conditions displaying a particular morphological phenotype of interest (e.g. high population variability), we identify genes and pathways correlated with this class distinction, thereby validating our previous studies, providing a means for studying determination of cell morphology, and generating new genes of interest for future study.

Overview of methods

Our overarching goal is to study differences in expression between treatment conditions showing different morphologies. This amounts to defining a class distinction to separate treatment conditions into groups of classes on the basis of morphology and then, subsequently, determining differential expression between these groups. We generate three different types of class distinctions corresponding to phenoclusters and variability analysis (in the last case, building off the results of Chapter 2). More specifically, we consider class distinctions defined by: phenoclusters versus control; and high/low morphological variability versus control and high versus low morphological variability.

For each class distinction, we select all treatment conditions from the *Drosophila* BG-2 morphology screen also present in a *Drosophila* microarray screen in a different cell line [63] that fall into the two groups dictated by the class distinction; we then determine genes that are

differentially expressed as well as pathways that are enriched between the two groups. As already noted, unlike for Chapters 2 and 3 where developing novel methods was necessary for successful analysis, the literature is rich for methods to determine differential expression or gene set enrichment between two unpaired groups. After normalizing the microarray data, we perform t-tests as well as significance analysis of microarrays (SAM) to determine differential expression, and we carry out gene set enrichment analysis (GSEA) to determine gene set enrichment. The gene sets we consider are all *Drosophila* gene sets in KEGG and GO as of June 2009.

Overview of results

Differential expression of single genes is essentially absent when using standard methods based on t-tests and correction for multiple hypothesis testing. Using a less stringent method (SAM), it is possible to identify single genes exhibiting moderate differential expression for some of the class distinctions under consideration. In many cases, individual genes that are identified by this analysis can be rationalized with the relevant class distinction. Gene set enrichment analysis, on the other hand, produces more substantive results. For example, for the class distinction defined by high versus low morphological variability, expression levels for the mTOR pathway are enriched for the high-variability treatment conditions.

It should be remarked that the *Drosophila* cell lines used for morphological and transcriptional data are different (BG-2 for morphology, S2R+ for transcription). This may help explain the lack of significant results when using t-tests, as signal strength is diminished when comparing alternate cell lines. On the other hand, because we do obtain some meaningful results even when

comparing different cell lines, we are encouraged to carry out further experiments to continue this line of research in future work – namely, to obtain microarray data for a screen using BG-2 cells.

Chapter 5: Conclusion

The three body chapters are each structured with an abstract, introduction, and sections for results, discussion, and materials and methods, references and figures. The concluding, fifth chapter again discusses the implications of each of the three main lines of work in this thesis, and puts all results in perspective as a full body of work. Limitations and future directions are also discussed.

References

1. Varmus, H. (2006). The new era in cancer research. Science *312*, 1162-1165.

2. Sporn, M. B. (1996). The war on cancer. Lancet *347*, 1377-1381.

3. Jaffe, A. B., and Hall, A. (2005). Rho GTPases: biochemistry and biology. Annu Rev Cell Dev Biol *21*, 247-269.

4. Ridley, A. J., Schwartz, M. A., Burridge, K., Firtel, R. A., Ginsberg, M. H., Borisy, G., Parsons, J. T., and Horwitz, A. R. (2003). Cell migration: integrating signals from front to back. Science *302*, 1704-1709.

5. Gomez del Pulgar, T., Benitah, S. A., Valeron, P. F., Espina, C., and Lacal, J. C. (2005). Rho GTPase expression in tumourigenesis: evidence for a significant link. Bioessays *27*, 602-613.

6. Heasman SJ, Ridley AJ. Mammalian Rho GTPases: new insights into their functions from in vivo studies. Nat Rev Mol Cell Biol. 2008 Sep ;9(9):690-701.

7. Bishop AL, Hall A. Rho GTPases and their effector proteins. Biochem J. 2000 Jun 1;348 Pt 2241-55.

8. Hall A. Rho GTPases and the Actin Cytoskeleton. Science. 1998 Jan 23;279(5350):509-514.

9. Bernards A. GAPs galore! A survey of putative Ras superfamily GTPase activating proteins in man and Drosophila. Biochim Biophys Acta. 2003 Mar 17;1603(2):47-82.

10. Amano M, Mukai H, Ono Y, Chihara K, Matsui T, Hamajima Y, Okawa K, Iwamatsu A, Kaibuchi K. Identification of a putative target for Rho as the serine-threonine kinase protein kinase N. Science. 1996 Feb 2;271(5249):648-50.

11. Watanabe N, Kato T, Fujita A, Ishizaki T, Narumiya S. Cooperation between mDia1 and ROCK in Rho-induced actin reorganization. Nat Cell Biol. 1999 Jul ;1(3):136-43.

12. Machesky LM, Insall RH. Scar1 and the related Wiskott–Aldrich syndrome protein, WASP, regulate the actin cytoskeleton through the Arp2/3 complex. Current Biology. 1998 Dec ;8(25):1347-1356.

13. Kobayashi K, Kuroda S, Fukata M, Nakamura T, Nagase T, Nomura N, Matsuura Y, Yoshida-Kubomura N, Iwamatsu A, Kaibuchi K. p140Sra-1 (specifically Rac1-associated protein) is a novel specific target for Rac1 small GTPase. J Biol Chem. 1998 Jan 2;273(1):291-5.

14. Tolias KF, Couvillon AD, Cantley LC, Carpenter CL. Characterization of a Rac1- and RhoGDI-Associated Lipid Kinase Signaling Complex. Mol. Cell. Biol. 1998 Feb 1;18(2):762-770.

15. Takagi Y, Ui-Tei K, Miyake T, Hirohashi S. Laminin-dependent integrin clustering with tyrosine-phosphorylated molecules in a Drosophila neuronal cell line. Neuroscience Letters. 1998 Mar 20;244(3):149-152.

16. Biyasheva A, Svitkina T, Kunda P, Baum B, Borisy G. Cascade pathway of filopodia formation downstream of SCAR. J Cell Sci. 2004 Feb 22;117(6):837-848.

17. Carpenter A.E., Jones T.R., Lamprecht M.R., Clarke C., Kang I.H., Friman O., Guertin D.A., Chang J. H., Lindquist R., Moffat J., Golland P., Sabatini D. M.: CellProfiler: image analysis software for identifying and quantifying cell phenotypes. Genome Biol. 7(10), R100 (2006)

18. Ohya Y., Sese J., Yukawa M., Sano F., Nakatani Y., Saito T.L., et al. High-dimensional and large-scale phenotyping of yeast mutants. Proc Natl Acad Sci U S A. 102(52), 19015-20 (2005)

19. Bakal C., Aach J., Church G., Perrimon N.: Quantitative Morphological Signatures Define Local Signaling Networks Regulating Cell Morphology. Science 316(5832), 1753-1756 (2007)

20. Berns K, Hijmans EM, Mullenders J, Brummelkamp TR, Velds A, Heimerikx M, Kerkhoven RM, Madiredjo M, Nijkamp W, Weigelt B, Agami R, Ge W, Cavet G, Linsley PS, Beijersbergen RL, Bernards R. A large-scale RNAi screen in human cells identifies new components of the p53 pathway. Nature. 2004 Mar 25;428(6981):431-437.

21. Boutros M, Kiger AA, Armknecht S, Kerr K, Hild M, Koch B, Haas SA, Consortium HFA, Paro R, Perrimon N. Genome-Wide RNAi Analysis of Growth and Viability in Drosophila Cells. Science. 2004 Feb 6;303(5659):832-835.

22. Loo L, Wu LF, Altschuler SJ. Image-based multivariate profiling of drug responses from single cells. Nat Methods. 2007 May ;4(5):445-53.

23. Paran Y, Ilan M, Kashman Y, Goldstein S, Liron Y, Geiger B, Kam Z. High-throughput screening of cellular features using high-resolution light-microscopy; application for profiling drug effects on cell adhesion. J Struct Biol. 2007 May ;158(2):233-43.

24. Rickardson L, Wickstrom M, Larsson R, Lovborg H. Image-Based Screening for the Identification of Novel Proteasome Inhibitors. J Biomol Screen. 2007 Mar 1;12(2):203-210.

25. Carpenter AE, Sabatini DM. Systematic genome-wide screens of gene function. Nat Rev Genet. 2004 Jan ;5(1):11-22.

26. Echeverri CJ, Perrimon N. High-throughput RNAi screening in cultured cells: a user's guide. Nat Rev Genet. 2006 May ;7(5):373-84.

27. Mitchison TJ. Small-molecule screening and profiling by using automated microscopy. Chembiochem. 2005 Jan ;6(1):33-9.

28. Pepperkok R, Ellenberg J. High-throughput fluorescence microscopy for systems biology. Nat Rev Mol Cell Biol. 2006 Sep ;7(9):690-6.

29. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet. 2000 May ;25(1):25-29.

30. Sacher R., Stergiou L., Pelkmans L.: Lessons from genetics: interpreting complex phenotypes in RNAi screens. Curr Opin Cell Biol. 20(4), 483-9 (2008)

31. Neumann B., Held M., Liebel U., Erfle H., Rogers P., Pepperkok R.: High-throughput RNAi screening by time-lapse imaging of live human cells. Nat Methods 3(5), 385-90 (2006)

32. Piano F., Schetter A.J., Morton D.G., Gunsalus K.C., Reinke V., Kim S.K., et al. Gene clustering based on RNAi phenotypes of ovary-enriched genes in C. elegans. Curr Biol. 12(22), 1959-64 (2002)

33. Gil J., Wu H., Wang B.Y.: Image analysis and morphometry in the diagnosis of breast cancer. Microsc Res Tech 59(2), 109-18 (2002)

34. Levy SF, Siegal ML. Network Hubs Buffer Environmental Variation in Saccharomyces cerevisiae. PLoS Biol. 2008 Nov 4;6(11):e264.

35. Slack MD, Martinez ED, Wu LF, Altschuler SJ. Characterizing heterogeneous cellular responses to perturbations. Proc. Natl. Acad. Sci. U.S.A. 2008 Dec 9;105(49):19306-19311.

36. Spencer SL, Gaudet S, Albeck JG, Burke JM, Sorger PK. Non-genetic origins of cell-to-cell variability in TRAIL-induced apoptosis. Nature. 2009 May 21;459(7245):428-432.

37. Blake W.J., Kaern M., Cantor C.R., Collins J.J.: Noise in eukaryotic gene expression. Nature 422(6932), 633-7 (2003)

38. Raser J.M., O'Shea E.K.: Control of stochasticity in eukaryotic gene expression. Science 304(5678), 1811-4 (2004)

39. Ozbudak E.M., Thattai M., Kurtser I., Grossman A.D., van Oudenaarden A.: Regulation of noise in the expression of a single gene. Nat Genet. 31(1), 69-73 (2002)

40. Pedraza JM, van Oudenaarden A. Noise propagation in gene networks. Science. 2005 Mar 25;307(5717):1965-9.

41. Friedman, N., Linial, M., Nachman, I., and Pe'er, D. (2000). Using Bayesian networks to analyze expression data. J Comput Biol 7, 601-620.

42. Friedman N. Inferring cellular networks using probabilistic graphical models. Science. 2004 Feb 6;303(5659):799-805.

43. Segal E, Friedman N, Kaminski N, Regev A, Koller D. From signatures to models: understanding cancer using microarrays. Nat Genet. 2005 Jun ;37 SupplS38-45.

44. Huang JC, Morris QD, Frey BJ. Bayesian inference of MicroRNA targets from sequence and expression data. J Comput Biol. 2007 Jun ;14(5):550-63.

45. Markowetz F, Spang R. Inferring cellular networks--a review. BMC Bioinformatics. 2007 ;8 Suppl 6S5.

46. Carter GW. Inferring network interactions within a cell. Brief Bioinform. 2005 Dec ;6(4):380-9.

47. Ma'ayan A. Network integration and graph analysis in mammalian molecular systems biology. IET Syst Biol. 2008 Sep ;2(5):206-21.

48. Cho KH, Choo SM, Jung SH, Kim JR, Choi HS, Kim J. Reverse engineering of gene regulatory networks. IET Syst Biol. 2007 May ;1(3):149-63.

49. Margolin AA, Califano A. Theory and limitations of genetic network inference from microarray data. Ann N Y Acad Sci. 2007 Dec ;111551-72.

50. Bhattacharjee, A., Richards, W. G., Staunton, J., Li, C., Monti, S., Vasa, P., Ladd, C., Beheshti, J., Bueno, R., Gillette, M., *et al.* (2001). Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. Proc Natl Acad Sci U S A *98*, 13790-13795.

51. Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., *et al.* (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science *286*, 531-537.

52. Barrett, T., and Edgar, R. (2006). Mining microarray data at NCBI's Gene Expression Omnibus (GEO)*. Methods Mol Biol *338*, 175-190.

53. Sachs K, Perez O, Pe'er D, Lauffenburger DA, Nolan GP. Causal Protein-Signaling Networks Derived from Multiparameter Single-Cell Data. Science. 2005 Apr 22;308(5721):523-529.

54. Pe'er D. Bayesian Network Analysis of Signaling Networks: A Primer. Sci. STKE. 2005 Apr 26;2005(281):pl4.

55. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science. 1999 Oct 15;286(5439):531-537.

56. Jiang N, Leach LJ, Hu X, Potokina E, Jia T, Druka A, Waugh R, Kearsey MJ, Luo ZW. Methods for evaluating gene expression from Affymetrix microarray datasets. BMC Bioinformatics. 2008 ;9284.

57. Hatfield GW, Hung S, Baldi P. Differential analysis of DNA microarray gene expression data. Mol. Microbiol. 2003 Feb ;47(4):871-877.

58. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proceedings of the National Academy of Sciences of the United States of America. 2001 Apr 24;98(9):5116-5121.

59. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences of the United States of America. 2005 Oct 25;102(43):15545-15550.

60. Mootha VK, Lindgren CM, Eriksson K, Subramanian A, Sihag S, Lehar J, Puigserver P, Carlsson E, Ridderstrale M, Laurila E, Houstis N, Daly MJ, Patterson N, Mesirov JP, Golub TR, Tamayo P, Spiegelman B, Lander ES, Hirschhorn JN, Altshuler D, Groop LC. PGC-1[alpha]-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nat Genet. 2003 Jul ;34(3):267-273.

61. Raj A, van Oudenaarden A. Nature, nurture, or chance: stochastic gene expression and its consequences. Cell. 2008 Oct 17;135(2):216-226.

62. Han JJ, Bertin N, Hao T, Goldberg DS, Berriz GF, Zhang LV, Dupuy D, Walhout AJM, Cusick ME, Roth FP, Vidal M. Evidence for dynamically organized modularity in the yeast protein-protein interaction network. Nature. 2004 Jul 1;430(6995):88-93.

63. Baym M, Bakal C, Perrimon N, Berger B.: High-Resolution Modeling of Cellular Signaling Networks. Proceedings of the 12th Annual International Conference on Research in Computational Molecular Biology (RECOMB 2008), LNBI 4955: 257-271, 2008.

Chapter 2:
Genetic Tuning of Morphological Variability in Cellular Processes

Abstract

A key challenge in systems biology is to analyze emerging high-throughput image-based data to understand how cellular phenotypes are genetically encoded. With the advent of technologies for acquisition of high-content imaging data, methods have been developed for quantifying cell shape, DNA morphology, and subcellular-localization of organelles or proteins. More subtle, however, is the problem of quantifying variability (i.e. noise) in cellular phenotypes, in effect, studying morphological variability itself as a cellular phenotype. Previous work has sought to quantify morphological variability, and has shown that knockout of network hubs results in decreased morphological noise. Other work has studied variability in occurrence of apoptosis. However, no previous work, to our knowledge, has utilized high-throughput image-based data to study variability in progression of cellular processes generally, or of cytokinesis specifically. Here, we first describe a robust, mathematically rigorous scoring procedure to quantify single-cell morphological variability on a population level. When applied to two high-content genetic screens in *S. cerevisiae* and *D. melanogaster*, our scoring procedure identifies roles for genes as modulators of morphological variability consistent with their known biological function, and consistent with previous studies on morphological variability. Further, by applying our scoring metric to sets of genes known to be involved in regulation of a specific cellular process (protrusion/adhesion formation and adhesion disassembly/cortical tension in fly; septin ring

formation in yeast), we identify genes which modulate morphological variability for that cellular process. We propose that some genes are either *suppressors* or *enhancers* of morphological noise for a process. Our results are supported by the known regulatory architecture for contractile ring formation in fly and septin ring formation in yeast. Overall, this study makes significant, new contributions to the young literature on morphological variability on the single-cell level in the context of genetic screens. We find that the effect of a genetic perturbation (knockout or overexpression of a gene) on morphological variability is explicable in many situations by the network architecture of the cellular process under consideration.

Introduction

Variability (i.e. noise) is an inherent property of signal transmission. Organisms have likely evolved to balance noise levels in signaling through regulatory networks in order to maximize phenotypic variability, without compromising the reliability of phenotypic responses [1]. For example, in order for metazoan cells to migrate across large distances towards guidance cues, cells make stochastic changes in morphology, such as formation of randomly oriented protrusions, thus increasing the reception potential of diffuse signals. After signal detection morphological changes become more predictable, promoting efficient migration towards the signal, and noise levels are essentially tuned down. The genes responsible for tuning the levels of noise in signaling pathways that regulate shape are unknown.

While the stochastic nature of some one-dimensional phenotypes such as the regulation of transcription have been previously quantified and explored [2-5], quantifying the stochasticity of cell shape regulation is more difficult because regulation of shape involves the control of

different cellular processes simultaneously. With the advent of image-based automated technologies and acquisition of high-throughput quantitative imaging data [6, 7], methods have recently been developed which attempt to use these technologies to quantify shape [8], DNA morphology [9], and subcellular-localization of organelles or proteins [10, 11], on a single-cell level. Analysis has commonly been performed by averaging single-cell results to derive mean scores for each genetic perturbation or by clustering such results [8, 12-14]. Recently, Levy et al. [15] quantified population variability by studying the variance of composite features in a genetic screen in yeast, and showed that knockout of genes with high network connectivity (i.e. network hubs) tends to increase observed morphological noise. However, Levy et al. do not analyze properties of genes which, when knocked out, decrease morphological noise, nor do they measure genetic contributions to morphological noise in specific cellular processes or study morphological variability in higher eukaryotes. Slack et al. [16] quantified population variability in a chemical screen in HeLa cells by viewing populations as mixtures of phenotypically distinct subpopulations and viewing chemical response as a redistribution of the relative subpopulation proportions. This approach was successful in classifying drugs according to mechanism of action. Consideration of genetic modulation of morphological noise in specific cellular processes, however, has been limited to apoptosis [17]. Quantifying noise in a non-lethal cellular process, where cell shape may change along multiple dimensions, is a much more complex problem than measuring presence/absence of cell death.

Here we introduce methods for measuring genetic contributions to morphological variability for specific cellular processes. We first develop a robust method for measuring population variability more generally, and apply this method to genetic screens in both yeast and fly. Subsequently, we apply our variability scoring procedure to study genetic contributions to

morphological variability in specific cellular processes (protrusion/adhesion formation and adhesion disassembly/cortical tension in fly; septin ring formation in yeast). We validate our results based on known gene functions and network architectures for the processes under consideration. We find that the effects of genetic perturbations on morphological variability are explicable in many situations by the network architecture of the cellular process under consideration.

Results

As a first step toward measuring genetic contributions to morphological variability within a specific cellular process, we introduce a scoring metric for measuring morphological variability for single-cell populations more generally (**Fig. 1**). Consider morphological data consisting of n genetic perturbations (called treatment conditions or TCs) across K feature dimensions, where the i^{th} TC, denoted TC_i, consists of c_i single cells (for the yeast screen [7], $n = 4787$ and $K = 101$; for fly [8], $n = 273$ and $K = 145$; **Materials and Methods** and **Supplementary Tables 1-3** contain further dataset descriptions). We performed normalization and dimensionality reduction of raw feature data to obtain reduced data of dimensionality k (**Materials and Methods** and **Supplementary Tables 4-5**); thus, data for TC_i was represented by a set of c_i points in reduced space. A variability v-score, v_i, and variability p-score, p_i, was calculated for each TC_i. Denoting the point set TC_i by $\{s_1, s_2, \ldots, s_{c_i}\}$, v_i is defined as the normalized average of the squared (Euclidean) distances of the $\{s_j\}$ from their center,

$$v_i = \frac{c_i}{c_i - 1}\overline{\|s_j - \bar{s}\|^2} = \frac{1}{c_i - 1}\sum_{j=1}^{c_i}\|s_j - \bar{s}\|^2.$$

An associated variability p-score, p_i was also calculated, where p_i measures the probability that a random sample from the set of all points over all TCs in reduced space is less spread out than the points comprising TC_i – equivalently, that a random sample of cells from the set of all single cells across all TCs is less heterogeneous than the cells comprising TC_i. In particular, if Y_n denotes the distribution of v-scores for sets of n points (drawn from the set of all points over all TCs in reduced space), then p_i is defined as $p_i = P(Y_{c_i} < v_i)$. We calculated variability p-scores for all TCs using bootstrapping. For one-dimensional data our metric reduces to the sample variance (see **Materials and Methods** for theoretical properties of the variability score). The quantities v_i and p_i are robust to method of dimensionality reduction and have small standard errors by jackknifing (**Supplementary Fig. 1, Supplementary Tables 6-8**). We also considered alternate approaches for measuring population variability; the method described here is superior for its simplicity, generality, and robustness (**Materials and Methods** and **Supplementary Fig. 2**).

Computation and analysis of variability scores for yeast and fly genetic screens

We applied our method to two high-content morphological screens in *S. cerevisiae* [7] and *D. melanogaster* [8], thus identifying genes in each organism which, when perturbed (either knocked-out or overexpressed) resulted in single-cell population of significantly high or low variability. Our results were consistent with known biology and previous studies [15], providing a means of validation for our scoring metric. Analysis of yeast and fly screens revealed that knockout or overexpression of certain genes can *increase* population variability, consistent with known function. We considered yeast TCs with the highest 1% of variability p-scores, and

found this gene set to be enriched for Gene Ontology categories [18] involved not only in morphogenesis, but also in chromosomal organization and DNA repair (**Materials and Methods** and **Table 1A**), consistent with the fact that disruption of these processes likely produces abnormally heterogeneous morphology. These results were consistent with those previously reported in yeast [15]. In the fly screen, a single TC, RNAi of *pbl*, resulted in a cell population with elevated variability at $p = 0.05$, after Bonferroni correction (**Table 1B**). Pbl is a RhoGEF known to regulate cytokinesis, adhesion formation, and mesenchymal development. This TC had large population variability likely because morphological processes regulated by *pbl* become noisy in its absence, reflecting pbl's importance in promoting orderly morphogenesis, particularly progression of cytokinesis. In yeast also, a single TC, knockout of CLA4, resulted in elevated variability at Bonferroni-corrected $p = 0.05$. CLA4 encodes an upstream activator of septin ring assembly which phosphorylates Cdc3p and Cdc10p, but whose exact function is unknown [19, 20]. We investigate the noise present in septin ring formation in greater detail below.

Knockout or overexpression of genes can also *decrease* population variability, again reflecting known molecular biology. For yeast, a total of 491 genes scored at $p < 10^{-8}$ (meaning that no bootstrapped samples had lower variability v-score), corresponding to Bonferroni-corrected $p < 5 \cdot 10^{-5}$. This set was enriched for genes involved in mitochondrial translation, perhaps due to impaired metabolism limiting ability to generate dynamic morphologies (**Table 2A**). For fly, 28 TCs had significantly reduced variability at Bonferroni-corrected $p = 0.05$. These represent single-cell populations with consistent morphologies, though these morphologies may be abnormal. For example, cells in the *apc2* knockout displayed unusually plentiful protrusions – but the cells in this population did so consistently. Likewise, Gef26- or armadillo-deficient cells

did not form protrusions, resulting in small, round cells. The set of 28 TCs was enriched for genes involved in protrusion/adhesion formation and lamellipodia formation by hypergeometric statistics (**Table 2B**). This finding reflects the fact that TCs unable to form these structures have abnormally homogeneous (small, round cell) morphology. Overall, when applied in a straightforward manner to image-based data from a genetic screen, our scoring procedure for population-level morphological variability identifies genes which, when knocked-out or overexpressed, result in populations displaying significantly increased or decreased variability. These results were consistent with those from previous studies and with known biological facts.

Morphological variability in cellular processes: Phenocluster analysis in fly

We next applied our metric to subsets of TCs in *Drosophila* previously implicated in control of particular morphological processes, thus identifying genes modulating morphological noise in those processes. Consider a TC defined by RNAi (an overexpression TC is symmetric). Genes that when inhibited by RNAi result in populations with high morphological variability are considered *suppressors of noise* (e.g. gene A in **Fig. 2**). Conversely, genes that when inhibited drive populations towards abnormal homogeneity are considered *enhancers of noise* (e.g. gene C in **Fig. 2**). To study suppressors and enhancers of morphological noise, we utilized TCs from the *Drosophila* screen thought to be involved in regulating a single morphological process, called phenoclusters [8], for example protrusion/adhesion formation. By comparing variability p-scores across TCs within a functionally-related cluster, we identified genes that modulate morphological noise for a particular morphological process. We calculated the variability p-scores for the three largest phenoclusters and identified TCs with significantly elevated or

depressed variability, relative to TCs in the same phenocluster (**Materials and Methods**). While there were no statistically significant results for lamellipodia formation (**Supplementary Table 9**), we now report results for the protrusion/adhesion formation and adhesion disassembly/cortical tension phenoclusters. Our findings regarding noise suppressors and enhancers agreed with known regulatory architecture of RhoGTPase signaling.

Several TCs displayed significantly decreased or increased population variability relative to the other TCs in the protrusion/adhesion formation phenocluster. In particular, *Goalpha65A_overex*, *Gef26*, *delRhoGEF3_const_overexp*, *Arf51F*, and *CG4448* all had significantly decreased variability. As discussed earlier, the set of *Drosophila* TCs that exhibited the lowest variability p-scores was enriched for TCs from the protrusion/adhesion formation and lamellipodia formation phenoclusters. Thus, it is particularly enlightening to ascertain which genes from these two phenoclusters actually have relatively *high* single-cell morphological variability. For the protrusion/adhesion formation phenocluster, no genes had significantly increased population variability after Bonferroni correction, though the two top scoring TCs, *RhoGEF3* and *RhoGAPp190*, did have marginally significant values (**Fig. 3B**, **Supplementary Table 10**). RhoGAPp190 is known to inhibit activity of the GTPase Rho1, while RhoGEF3 promotes Rho1 activity [8, 21-24]. Furthermore, these proteins work immediately upstream of Rho1 in their signal transduction pathways. Extensive work in measuring single-gene transcription has shown that perturbation of expression of genes with upstream products causes increased noise in the expression of downstream targets [2, 25]. The fact that RhoGEF3 and RhoGAPp190 knockdowns result in the greatest observed morphological noise among all TCs in the protrusion/adhesion formation phenocluster supports the validity of our methods for measuring variability. We propose that *RhoGEF3*, *RhoGAPp190*, and *Goalpha65A* are suppressors of

noise, while *Gef26*, *Arf51F*, and *CG4448* are enhancers of noise for the process of protrusion/adhesion formation (**Fig. 3A**, **Supplementary Figs. 3-5**).

For adhesion disassembly/cortical tension, again several TCs displayed significantly reduced variability (**Fig. 3D**, **Supplementary Table 11**). The TC with the smallest p value was RacGAP50C, which plays an important role in regulation of contractile ring formation in cytokinesis. Furthermore, pbl knockout resulted in the largest p value in this phenocluster; this variability p-score represented significantly increased variability for this morphological process (after Bonferroni correction). The roles of RacGAP50C as an enhancer of noise and of pbl as a suppressor of noise (**Fig. 3C**, **Supplementary Fig. 6**) are consistent with the spatiotemporal dynamics of the protein network regulating contractile ring formation. It is known that both RacGAP50C (along with its binding partner, the Pavarotti protein) and pbl are required for proper contractile ring formation and progression of cytokinesis [26, 27]. More specifically, the RacGAP50C protein is required for furrow formation at the beginning of cytokinesis, and absence of the protein causes an early failure in cytokinesis [28]. As discussed above, absence of pbl also disrupts cytokinesis, but its role is more extensive; it is involved in vesicle trafficking and actin recruitment to the furrow. Both RacGAP50C and pbl are required for cytokinesis progression, but knockout of one resulted in highly variable, aberrant morphology, while knockout of the other resulted in consistent (though still abnormal) cell-to-cell morphology.

In sum, by applying our variability scoring procedure to subsets of TCs in *Drosophila* previously implicated in control of particular morphological processes, we were able to identify genes modulating variability in these processes. These findings were consistent with the known regulatory architecture of the processes under consideration.

Morphological variability in cellular processes: Septin ring formation in yeast

We further validated our methods by studying septin ring formation using the yeast screen, obtaining variability results consistent with known network architecture and basic concepts of signal processing. We first consider the observed morphological noise present in TCs for genes whose role in septin ring formation is well-characterized, and show that the level of noise is consistent with the known biological roles of these genes as well as their place in the regulatory architecture. Subsequently, we study morphological noise present in TCs for genes thought to play a role in septin ring regulation, but whose function is not well-characterized. We use our methods to propose new roles for several of these genes. Specifically, we formulated seven hypotheses regarding morphological variability which follow from the known regulatory structure of septin ring formation. With one exception, all of these hypotheses were borne out directly by variability p-score calculations.

The septins, which in *S. cerevisiae* consist of CDC3, CDC10, CDC11, CDC12, and SHS1 (SEP7), are a conserved family of proteins that form a scaffold for localization of other proteins involved in bud site selection and cytokinesis [29-31]. It is thought that septin ring formation consists of two distinct steps, recruitment and assembly [32]. Recruitment depends on activity of Cdc42p, which is promoted its GEF, Cdc24p. Subsequently, assembly of the septin ring is promoted by several proteins acting in parallel pathways including Cla4p, Gin4p, Nap1p, Bni5p, and Elm1p, as well as the three Cdc42p GAPs Bem3p, Rga1p, and Rga2p [19-20, 32-35]. The exact mechanisms of the non-GAP regulatory proteins have been explored over the past decade, but have not been established definitively.

In the process of bud formation in wild-type cells, there is a transition from apical bud growth to isotropic growth. This transition is governed by activity of Clb-Cdc28p, which is governed in

turn by multiple regulators. In particular, Swe1p inhibits activity of Cdc28p via phosphorylation of Tyr 19. Regulation of Swe1p occurs in turn via Hsl1p and Hsl7p: the septin scaffold recruits Hsl1p, which interacts with Hsl7p-Swe1p, leading to the rapid degradation of Swe1p. When Swe1p is active, the cell is arrested in G2; following degradation, the cell proceeds through the G2/M transition. Thus, overexpression of Swe1p causes G2 arrest and formation of elongated buds of variable morphology [36].

The following hypotheses derive from this regulatory architecture for septin recruitment and assembly. 1) Knockout of CDC42 or CDC24 should either be lethal or cause abnormal and highly variable morphology due to improper septin recruitment and subsequent defects in bud formation and cytokinesis; 2) Knockout of any of the five septins should likewise either be lethal or cause abnormal, variable morphology due to defective septin assembly and sequellae (**Fig. 4B** and **Supplementary Fig. 7**); 3) Knockout of CLA4, GIN4, NAP1, BNI5, or ELM1 should likewise result in abnormal, variable morphology due to improper septin assembly and sequellae (**Fig. 4A** and **Supplementary Fig. 8**); 4) Knockout of HSL1 or HSL7 should result in abnormal, variable morphology due to the resultant constitutive activity of Swe1p leading to formation of elongated buds of variable size (**Fig. 4C** and **Supplementary Fig. 9**); 5) Knockout of SWE1 should not result in cells of variable morphology, since these cells should pass through the G2/M transition rapidly leading to small buds of relatively consistent morphology (**Fig. 4C** and **Supplementary Fig. 9**); 6) Knockout of any one of the three Cdc42p GAPs should result in cells of relatively consistent morphology [32] (**Fig. 4D**); 7) Knockout of CDC28 should either be lethal or result in cells of highly variable (bud) morphology, since the transition from apical to isotropic growth would not occur.

All but one of the seven hypotheses was borne out by the variability score data (**Fig. 4D** and **Supplementary Table 12**). The one exception was the hypothesis regarding CLA4, GIN4, NAP1, BNI5, and ELM1. In particular, we found that knockout of CLA4 or ELM1 resulted in highly variable single-cell populations, while knockout of GIN4, NAP1, or BNI5 did not. In the case of BNI5, partially this was due to the image segmentation of [7], wherein highly clumped cells were typically not segmented. When clumping is promoted by the TC gene knockout, there may be a bias towards segmenting cells of relatively normal morphology. In particular, for BNI5, cells with elongated buds were more likely to clump together and/or overlap, meaning that these cells were less likely to be segmented. However, this did not occur for NAP1 or GIN4 cells. Excluding BNI5, how then do we account for the fact that CLA4 and ELM1 knockout TCs have high single-cell variability, whereas NAP1 and GIN4 knockouts do not? One possibility is temporal regulation of parallel pathways affecting septin formation; perturbations of proteins acting earlier in the regulatory process is likely to cause more downstream morphological variability than perturbations of late-acting proteins. This explanation is supported by the fact that cla4p acts early in budding [20] and the fact that CLA4 displayed the largest variability p-score. A second, related possibility is that two or more of these proteins are acting in a linear pathway, with knockout of downstream proteins resulting in less morphological variability. Putting the underlying mechanism aside, the basic fact is that knockout of different genes regulating septin ring formation results in huge variations in population-level morphological variability. Looking at this from the other side, this means that activity of certain proteins, namely cla4p or elm1p, has the effect of homogenizing cell-to-cell morphology, whereas activity of other proteins, namely nap1p or gin4p, has the contrary effect of promoting cell-to-cell heterogeneity. Thus, we propose that CLA4 and ELM1 are suppressors of noise and that GIN4

and NAP1 are enhancers of noise for the process of septin ring assembly. Overall, the fact that multiple hypotheses regarding population variability for TCs related to regulation of septin ring formation were supported by the variability data provides further validation for the methods introduced here to measure morphological variability in cellular processes.

Discussion

We developed methods to probe the genetic basis of morphological stochasticity. Because ours are the first methods to quantitatively study the genetic regulation of morphological variability in cellular processes, we validated our results using known functional properties of genes in our datasets and regulatory architecture of relevant cellular processes. As additional sources of morphological data (different cell lines, different organisms) become available, we expect these methods to serve to quantify morphological variability of single-cell populations and genetic modulation of morphological stability and variation in cellular processes.

Alternative methods for measuring morphological variability

Here we introduced a scoring procedure for measuring morphological variability of a population of cells, in the situation where we have image-based data for each cell comprising the population. We showed by extensive testing that our method is robust. Furthermore, the basis for the metric is relatively simple: a multi-dimensional analog of one-dimensional variance, applied to data that has been normalized by taking z-scores in each raw dimension and then

reduced in dimensionality by using PCA. The benefit of applying a relatively simple methodology is increased confidence in the interpretation of results.

A competing method of measuring morphological variability [15] is less intuitive. This method performed dimensionality reduction by using a clustering procedure to pick a large number of representative features. Subsequently, variances (normalized, to account for dependence on mean) were computed for the 70 representative features, and the average of the top 35 features was used as the variability score. The researchers noted that their method was not robust when a relatively small number of top features were used for the average, and suggested therefore that "genes that cause a high variance in only a few phenotypes are different from those that cause high variance globally." Their method was effective at identifying genes which, when knocked out, result in highly variable morphology in a great many feature dimensions. In contrast, our method is effective at studying more subtle changes in morphological variability in cellular processes. It is also effective at identifying genes which, when knocked out, result in decreased variability. We are less interested, here, in knockouts which cause global disturbances, i.e. which extensively break cellular morphology.

As an additional note, normalization prevents arbitrariness in units for raw feature measurement from adversely affecting the variability score. For example, consider the difference it makes if we measure cell perimeter (a typical example of a raw geometric feature) in millimeters versus centimeters. In the latter case, all measurements are multiplied by a factor of 10. Thus, the mean for this feature dimension increases by a factor of 10, and the variance increases by 100. In fact, there is complete arbitrariness in what units are used for all raw feature measurements, so it is inappropriate to compare variances directly. Furthermore, it is not desirable to consider the ratio of standard deviation to the mean as is often done for transcriptional data [25] (this ratio is

invariant to unit scaling) because morphological traits are not, in general, interpretable as frequencies ("counts"); accordingly, they do not have well-defined means (for example, there is no reason, *a priori*, that the morphological feature measuring eccentricity of an (elliptical) cell should be a number between 0 and 1 rather than a number between, say, -1 and 1). To account for this arbitrariness, we normalize all raw features so that they have mean 0 and variance 1. Subsequently, by considering differences in spread between the set of points in a particular TC in normalized coordinates, on the one hand, and a randomly selected set of points from the entire screen, on the other, we quantify variability in the TC population. In their paper, Levy et al. argue against PCA as a method of dimensionality reduction because "loadings of the initial data may be negative. A high value in a principal component may represent a high or low variance in the underlying phenotypes." While this is a true statement, it misses the point, and does not argue against using PCA as a means of dimensionality reduction if applied carefully. Indeed, the scoring procedure that we use considers the *variance* of PCA-based components, not their *values*. Furthermore, as noted above, it is inappropriate to put credence in high/low variance of raw feature values because of arbitrariness of units. What needs to be done, instead, is to uncover the structure of the data by fist normalizing all raw features then removing redundant features, and finally to measure variance/spread for each TC in this reduced space.

The role of network architecture in modulating morphological variability

In the course of our analysis, we have shown that population-level morphological variability reflects the architecture of regulatory networks. Our results extend the finding of Levy et al. [15] that knockout of network hubs tends to increase morphological variability. Here, we considered

more intricate network architectures associated with regulation of complex cellular processes. Indeed, work in measuring single-gene transcription has shown that perturbation of expression of genes with upstream products causes increased noise in the expression of downstream targets [2, 25]. We demonstrated repeatedly that perturbation of genes acting upstream in signaling pathways tends to increase morphological noise in the process mediated by the pathway to a greater extent than perturbation of genes acting further downstream in the pathway. For example, in the case of septin ring assembly, knockout of HSL1 or HSL7 increases morphological variability to a far greater extent than knockout of SWE1, and perturbation of the upstream activators CLA4 and ELM1 resulted in increased variability.

We resist making the naïve teleological conclusion, however, that these genes function to increase or decrease noise. We do speculate, as have others in the case of gene expression [2-5, 25], that regulatory networks evolve to allow for subtle tuning of morphological variability. Certain genes, which occupy key positions in the network's spatiotemporal architecture, can increase/decrease morphological noise in the cellular process by virtue of increased/decreased activation. These genes generally also have direct roles in regulating the process at hand, so they are not solely suppressors or enhancers of morphological noise. Rather, they are *also* suppressors or enhancers of morphological noise, in addition to their other functions.

The fact that perturbation of upstream gene products tends to increase variability relative to downstream products can potentially be used as an informative signal in inference procedures for signaling pathways. More specifically, an effective framework for pathway inference based on morphological data would make use of the fact that knockouts of upstream proteins tends to increase observed population variability p-scores while maintaining relatively similar mean morphology. Such a framework would need to be combined with other data sources (e.g.

transcriptional) to obtain meaningful results {cite Baym}. But the key to using morphological data to obtain directionality in signaling networks is to use measurement of population-level variability, as this capitalizes on the noise propagation properties of signaling pathways, as demonstrated repeatedly in this paper.

A complication of using this property as an informative signal in pathway inference, of course, is that morphological variability following perturbation of a gene is influenced by other spatiotemporal properties of regulatory networks. For example, in the case of contractile ring formation in fly, we found that knockout of RacGAP50C decreases morphological noise while knockout of pbl increases morphological noise, even though both genes act upstream of Rac1. This reflects the temporal dynamics of cytokinesis. Namely, RacGAP50C acts early in cytokinesis by regulating furrow formation, whereas pbl acts later. Morphological noise is decreased by RacGAP50C knockout because cells are locked in an abnormal configuration that tends to be highly consistent from cell to cell, whereas morphological noise in increased by pbl knockout, as cells are abnormal but variable. Here, the difference in morphological variability depends on temporal regulation of cytokinesis progression, not on directionality in a signaling pathway. Further, in the case of septin ring formation, we found that perturbation of the upstream activators CLA4 or ELM1 resulted in increased variability, but perturbation of NAP1 or GIN4 did not. According to the hypothesis regarding the effect of directionality on variability, we would expect increased variability in all cases. This finding likely reflects the complex dynamics of septin ring regulation, similar to contractile ring formation in *Drosophila*.

Overall, our work represents an important step in probing genetic contributions to morphological variability in cellular processes, and connecting the modulation of variability to network structure. Further work will study the effects of different regulatory structures on modulation of

variability by detailed study of other cellular processes. At present, this is limited by the fact that certain cellular processes tend to produce more much dramatic morphological changes than others. Basic clustering of morphological signatures is able to group genes for several processes [8], but our methods do not find significant results in all cases (namely, we were unable to identify modulators of noise for lamellipodia formation in fly). Further refinement of our methods, as well as new genetic screens to obtain image-based data, will be necessary to increase the sensitivity for less dramatically modulated cellular processes.

Materials and Methods

Morphological datasets

As described in [8], TCs were prepared in the *Drosophila* DM-BG2 (referred to as BG-2) cell line using either dsRNA or overexpression constructs. The screen consisted of 249 distinct genetic perturbations, with several replicates, for a total of 273 TCs. The 249 TCs correspond to 45 dsRNAs targeting Rho GTPases, GEFs, and GAPs, 20 overexpression constructs and 173 dsRNAs chosen randomly from a set of genes implicated in cytoskeletal organization, and overexpression of SIF (a *Drosophila* RhoGEF) in combination with several randomly selected dsRNAs. The full set of TCs is listed in **Supplementary Table 1**. Cell segmentation was performed using the custom CellSegmenter Software. Cells were stochastically labeled with GFP to facilitate image segmentation. For each cell, 145 geometric features and 9 status features were extracted in a semi-automated fashion (see [8] for details). The full list of geometric features is given in **Supplementary Table 2**.

For yeast, as described in [7], the genetic screen consisted of 4787 distinct genetic perturbations, essentially consisting of all non-lethal single gene knockouts. Cell segmentation and image analysis was performed using custom-built software. For each cell, a total of 158 geometric features were measured, of which we used 101 (the features that were not used included, for example, coordinates of the cell center; these were used in [7] to compute derived features, which we do not make use of here). Some cells were missing data for features, in which case we used the mean feature value across single cells in the same TC (for which the data was available). This has the effect of maintaining, for each TC, the mean for each feature as well as the squared deviation for each feature. See **Supplementary Table 3** for a list of geometric features measured by [7] and those used in our variability analysis.

Data normalization and dimensionality reduction

For both the *Drosophila* and yeast datasets, single-cell data was normalized across each raw feature dimension to that the full set had mean 0 and variance 1 for each raw feature. Transforming the data so that each raw feature has equal variance is required because our goal is to measure differences in variability for different TCs, i.e. different subsets of the full set of single cells. Since we will later use subset variances in our variability score calculations, normalization of the raw features must be performed to avoid inappropriately weighting some features over others (for example, because of arbitrary differences in unit measurements) in the variability calculations. The transformation so that each raw feature has mean 0 is necessary prior to calculating principal components; on the other hand, it is not strictly necessary to

transform the data so that each raw feature has variance 1 (the variances simply need to be equal, and the principal component calculations will be the same either way).

Following data normalization, dimensionality reduction was performed by computing principal components for the full set of single-cell data, and then projecting each data point onto the first three principal components. This has the net effect of reducing dependencies between raw features, and is necessary to avoid inappropriately weighting our variability scores towards particular morphological feature classes that are overrepresented in the set of raw features (for example, redundant measurements of nucleus shape).

Calculation of principal components was performed in Matlab for the *Drosophila* data, and was performed by writing custom Java code for the yeast data. The latter approach was necessary because the yeast data set contains approximately 1.9M single cells versus approximately 12K for the *Drosophila* set. Therefore, it was impractical to compute the principal components for the yeast data by calculating the covariance matrix, which is the typical derivation of PCA. Instead, an iterative method based on Expectation Maximization was used. For this method, each raw feature must have mean 0 across the full set of single-cell data, as was accomplished in data normalization (see above). The algorithm proceeds by initializing a random vector q_0 and updating it by performing a number of iterations through the full set of single-cell data. The algorithm terminates when the vector changes by $< 10^{-4}$ in the max norm, i.e. when $|q_i - q_{i+1}| < 10^{-4}$. The update of q_i to q_{i+1} consists of projecting each single-cell vector onto q_i, and summing these projections over the entire dataset. This sum is then normalized to have magnitude 1 and defined to be q_{i+1}.

Running this algorithm has the effect of calculating the first principal component for the dataset. Subsequently, we subtract the first component from each single-cell vector (i.e. for each vector, we project onto this component and subtract this projection from the vector). The second principal component can then be calculated by re-running the algorithm on the reduced data, and so on for the next principal components. The first three PCs for the *Drosophila* and yeast datasets are reported here (**Supplementary Tables 4-5**).

Theoretical properties of variability v- and p-scores

Recall that the data for TC is represented by a set of n points, $\{s_1, s_2, \ldots, s_n\}$, of dimension k in reduced feature space. Let S denote the corresponding set of points in reduced feature space. Then the variability v-score, v, is defined to be the average of the squared (Euclidean) distances of each point in TC from their center of mass, weighted by a normalization constant ($\frac{n}{n-1}$),

$$v = \frac{n}{n-1}\overline{\|s_j - \bar{s}\|^2} = \frac{1}{n-1}\sum_{j=1}^{n}\|s_j - \bar{s}\|^2.$$

Note that this statistic is the one-dimensional analog of the variance obtained by applying the Euclidean norm to the k-dimensional points, $\{s_1, s_2, \ldots, s_n\}$. Let Y_n denote the distribution of this statistic calculated on n points (drawn from S). We now consider theoretical properties of the distribution Y_n.

In the case where $k = 1$, e.g. where we use just the first principal component in order to reduce each single cell to a one-dimensional morphology score, then the distribution Y_n reduces to the

chi-square distribution with $n-1$ degrees of freedom, on the assumption that S is normally distributed. To see this, note that in the case $k = 1$, we have

$$v = \frac{1}{n-1}\sum_{j=1}^{n}\|s_j - \bar{s}\|^2 = \frac{1}{n-1}\sum_{j=1}^{n}(s_j - \bar{s})^2 = var(\{s_j\}),$$

which is to say, the variability v-score, v, reduces to the usual one-dimensional variance of the points $\{s_j\}$. Assuming that the underlying population S is normally distributed with variance 1, then the sample variance is known to follow the chi-square distribution with $n-1$ degrees of freedom, so $v \sim \chi^2_{n-1}$.

Now consider the case where $k > 1$, which is more realistic. Write each point s_j in k-dimensional coordinates; that is, $s_j = (s_{j1}, s_{j2}, \ldots, s_{jk})$, and let $\bar{s} = (\bar{s_1}, \bar{s_2}, \ldots \bar{s_k})$. Then, by definition,

$$v = \frac{1}{n-1}\sum_{j=1}^{n}\|(s_{j1}, s_{j2}, \ldots, s_{jk}) - (\bar{s_1}, \bar{s_2}, \ldots \bar{s_k})\|^2 = \frac{1}{n-1}\sum_{j=1}^{n}\sum_{l=1}^{k}(s_{jl} - \bar{s_l})^2.$$

Exchanging the order of summation and rearranging yields

$$v = \sum_{l=1}^{k}\left[\frac{1}{n-1}\sum_{j=1}^{n}(s_{jl} - \bar{s_l})^2\right] = \sum_{l=1}^{k} t_l.$$

Assume that S is distributed as a multivariate normal where the l^{th} dimension of S has mean 0 and variance w_l. We cannot assume that each dimension has the same variance; typically, since we use PCA to perform the dimensionality reduction from raw feature data to obtain S, the l^{th} dimension has variance equal to the variance contribution of the l^{th} principal component. Now, the quantity, t_l, in the summation above is the sample variance of one-dimensional numbers

drawn from the l^{th} dimension of S. Therefore, $t_l \sim w_l \cdot \chi^2_{n-1}$. This means that we v can be thought of being the sum of k (scaled) independent chi-square random variables. Knowing the variance contributions of the PCs tells us what these scaling are, and allows us readily to sample from these random variables.

The upshot of this derivation is that bootstrapping (see below) to determine the probability distribution Y_n can be simplified if S is distributed as a multivariate normal. Rather than repeatedly drawing n points from the underlying population distribution S, and computing their variability v-score, we can instead sample from the scaled chi-square distributions and use these to compute variability v-scores. This would be particularly helpful for the yeast dataset, where the population S is very large and a naïve approach to bootstrapping requires running through the entire dataset, S, for each bootstrapping iteration on disk, which is prohibitively slow. With appropriate supercomputing resources, the naïve implementation could execute in non-prohibitive time, but inasmuch as still larger datasets will likely become available in the future, it is worthwhile to develop computationally efficient methods.

On the other hand, the downside of using the theoretical chi-square distribution is that it is only valid if the underling population S is distributed as a multivariate normal. In fact, the raw feature data is most certainly not normally distributed, and it is not clear that transformation to PC-based coordinates produces normality. If neural networks are used to perform the dimensionality reduction of the raw data, then the reduced data, S, is not normally distributed. As a result, the most reliable approach to accurately determining the distribution Y_n and the variability p-scores is to perform bootstrapping by faithfully re-sampling from S and computing the test statistic. In the case of the yeast data, this requires writing bootstrapping code that is better than the naïve method; the trick is to compute and store the indices of multiple random samples of size n, and

compute the test statistic for all of these samples in a single pass through S on disk. This allows for both efficient and correct computation of Y_n.

As an additional theoretical note, our variability metric is related to the Mahalanobis distance [37]. The Mahalanobis distance is typically used to measure the distance of a test point, x, from a set of points $\{x_i\}$. The Mahalanobis distance is defined to be a weighted Euclidean distance between the point x and the center of mass, u, of the set $\{x_i\}$. Namely, the squared Mahalanobis distance is given by $(x-u)^T S^{-1}(x-u)$, where S is the covariance matrix of the $\{x_i\}$. The normalization by the covariance matrix distance accounts for the non-spherical nature of the set $\{x_i\}$, and is necessary for the goal of measuring distance of the test point, x, from the set $\{x_i\}$. However, in our case the goal is to measure the spread of the point set. Furthermore, since we have already transformed to coordinates in principal components, it is necessary to use a non-weighted Euclidean distance function in order to include the effects of differences in variance contributions of the principal components.

Bootstrapping

We wrote code in Matlab (for *Drosophila*) and in Java (for yeast) to compute the variability p-score corresponding to each variability v-score using bootstrapping using 10^5 iterations for *Drosophila* and 10^8 iterations for yeast. The number of iterations was selected to achieve sufficient resolution for statistical significance following Bonferroni correction for testing multiple hypotheses.

In greater detail, recall that the variability p-score corresponding to a variability v-score, v, for a TC consisting of n single cells is defined to be $p = P(Y_n < v)$, where Y_n denotes the distribution of the average squared distance of n points (drawn randomly from the entire set of single cells across all TCs) from their center of mass, weighted by the fraction $\frac{n-1}{n}$. Thus, for each n, we computed the variability v-score for either 10^5 (*Drosophila*) or 10^8 (yeast) randomly selected sets of n single cells from the full set of single cells across all TCs. The fraction of these scores that fell below v is the bootstrap calculated value for $p = P(Y_n < v)$.

Robustness to method of dimensionality reduction: Number of PC dimensions, neural networks

For our final calculations for both *Drosophila* and yeast, we used a dimensionality reduction based on PCA, using the first three principal components. As noted previously, there is some arbitrariness in the selection of three PCs instead of some other number. Also, it is possible to perform the dimensionality reduction using other methods besides PCA; for example, one can use the neural network classifiers developed in [8] to define reduced feature space.

We show here, however, that despite the arbitrariness in selection of number of PCs (and the choice of using PCA versus neural network classifiers) as the method of dimensionality reduction, the ordering of variability p-scores obtained in the end is remarkably invariant to the method of dimensionality reduction. To show this, we use the *Drosophila* data, and compute variability p-scores using bootstrapping with 10^3 iterations for reduced data using the either the first 1, 2, 3, 5, or 10 PCs, and also using the best 7 neural network classifiers (see [8] for discussion of relative efficacy of the neural network classifiers), for a total of six different

methods of dimensionality reduction. For each of the six methods, we considered the set of TCs having variability p-score $< 10^{-3}$ (i.e. TCs for which none of the bootstrapped samples had a smaller variability v-score than the TC itself). We then computed the probability of observed overlap of each pair of these six sets using the hypergeometric cdf. In addition, for each of the six methods of dimensionality reduction, we considered the set of 10 TCs with the largest variability p-scores. We again computed the probability of observed overlap of each pair of these six sets.

In nearly all cases, the probability of pairwise overlap is $< 10^{-9}$, which represents highly significant overlap (**Supplementary Tables 6-7**). The one exception is the overlap between the set of 10 top-scoring TCs when neural networks are used, which deviates to a relatively large extent from the set of 10 top-scoring TCs for each PC-based dimensionality reduction. The likely reason for this deviation stems from the fact that the neural network classifiers were not systematically constructed to incorporate information from all raw features, but rather to classify several specific, archetypal shapes. Thus, using the neural network classifiers as a means of dimensionality reduction is likely to result in lost information and amounts to an incomplete approach for measuring variability.

Robustness to data collection: Jackknifing

We used jackknifing to determine the standard error for each variability p-score in the *Drosophila* screen. Because these values were small compared to the variability p-scores themselves, particularly for TCs with extreme high or low variability p-scores, we gained confidence in the robustness of our methods for measuring population variability.

For each TC, we performed the following procedure. Remove a single-cell from the TC at random and re-compute the variability v-score for this new set. This procedure was repeated for each single cell for each TC in the *Drosophila* screen. Considering the variance of this set of re-computed variability v-scores allows us to calculate a standard error for each variability v-score that represents the effect of data collection errors. Subsequently, we used bootstrapping (with 10^5 iterations) to determine the standard error for each TC's variability p-score that corresponds to the standard error for that TC's variability v-score (**Supplementary Fig. 1** and **Supplementary Table 8**). Standard errors tend to be relatively large for TCs scoring in the mid-range for variability p-scores. For TCs with extreme high or low variability p-scores, the standard errors are small, in the sense that statistical significance at the Bonferroni-corrected value of $p = .05$ is maintained if two standard errors are added or subtracted from the observed variability p-scores.

Alternate methods for measuring population variability

We explored alternate approaches to measuring morphological variability of a single-cell population. In one approach, we developed a technique to quantify the morphological variability in different TCs by assigning a graph-based noise signature to each TC, called a Feature Graph (FG). We equate morphological variability/stochasticity in a population to the levels of noise in the population. Simplistically, populations of cells with very similar shapes have highly connected FGs, whereas morphologically heterogeneous populations have poorly connected FGs.

A FG is defined for each TC by the following procedure (depicted in **Supplementary Fig. 2**) to establish a graph-based noise signature for each TC. Each FG has the same set of nodes, which correspond to the set of features; for notational ease, these are also denoted by $\{F_1, F_2, \ldots, F_k\}$. For a given TC, the Pearson correlation matrix of linear correlations between pairs of features in the $c_i \times k$ matrix of feature scores is computed. Undirected edges are drawn between two nodes if the magnitude of the linear correlation between the two corresponding features across the population of cells exceeds a certain threshold, α. More specifically, an edge is drawn between the vertices F_l and F_m of the FG if the linear correlation $corr_{i,l,m}$ between features F_l and F_m is larger than α or smaller than –α.

We can then use FGs in order to compute a score for population variability. Specifically, we calculate the L_1 norm of the Pearson correlation matrix for each TC_i and normalize it by twice the median of the set of such norms. The L_1 norm is used because it provides an effective, computationally efficient means of measuring the overall variability present in a population of cells by combining the Pearson correlations between all pairs of morphological features. To understand intuitively why the L_1 norm is appropriate, consider the effect of adding a new feature, F_{k+1}. If this new feature is highly correlated with many of the previously measured features, then the L_1 norm will increase by a large amount; if the new feature is not highly correlated with the other features, then the norm will increase by a relatively small amount. Thus, the norm measures the extent to which the full set of morphological features are correlated across the TC. Other norms on the correlation matrix could be used to measure population variability, but comparable results are obtained. The normalization is performed in order to provide a framework for comparing the population variability across different TCs, particularly

to identify TCs having relatively low or high morphological variability. Using PV_i to denote the corresponding value for TC_i, we have:

$$PV_i = 1 - \frac{\|TC_i\|_1}{2 \cdot \mu_{\frac{1}{2}}(\|TC_i\|_1)_i},$$

where $\|TC_i\|_1 = \sum_{i=1}^{10} \sum_{m=i+1}^{10} |corr_{i,m}|$ and $\mu_{\frac{1}{2}}$ denotes the median.

For this calculation, we take advantage of the full Pearson correlation matrix without using an edge-determination threshold for the FG. Based on this formulation, a low score for PV_i means relatively low population variability of TC_i, while a high score is a reflection of high population variability We also computed standard errors for population variability for each TC in like manner to variability p-scores, as explained earlier.

FGs do provide some measure of population variability. The issue is that this measure is contrived in comparison to the much more straightforward approach of variability v-scores and variability p-scores defined in the main text. The variability p-score method is more robust and interpretable, and also satisfies the principle of maintaining straightforward, rigorous mathematics in computational biology.

Enrichment statistics

Gene Ontology enrichment statistics were calculated for yeast TCs in the top 1% of variability p-scores (**Table 1A**), and yeast TCs having variability p-score less than 10^{-8} (i.e. TCs for which no bootstrapped samples of equal cell number had larger variability v-score; see **Table 2A**). The

background set used for enrichment calculation was the entire set of 4787 genes in the screen. Enrichment significance was calculated using the hypergeometric cdf to determine the probability of enrichment for each GO category. Bonferroni correction was performed to account for testing multiple hypotheses.

Enrichment statistics were also examined for the set of *Drosophila* TCs having significantly reduced population variability relative to random (**Table 2B**). There was no significant GO enrichment for this set of 28 TCs, likely for the reason that the background set of 273 *Drosophila* TCs is highly enriched for genes involved in cytoskeletal regulation, thus raising the standard for GO enrichment (particularly after Bonferroni correction). However, we observed that the set of *Drosophila* TCs in the top 1% of variability p-scores appeared to be enriched for TCs classified into the phenoclusters for protrusion/adhesion formation and lamellipodia formation.

We used the hypergeometric cdf to show that this enrichment was in fact statistically significant. Out of the 28 TCs with significantly reduced population variability, 19 are involved in either lamellipodia formation or protrusion/adhesion formation, while 63 of the original 273 TCs belong to the union of these two categories. The probability of this overlap occurring by chance is $< 9 \cdot 10^{-9}$, as given by the hypergeometric cdf, indicating significant enrichment. (This is true even after correcting for multiple hypotheses, as there 820 pairwise combinations of the 41 phenoclusters and, to be extremely conservative, 273 choices for where to draw a cutoff for inclusion in a group of lowest-scoring TCs, meaning that $p < \frac{.05}{820 \cdot 273} = 2 \cdot 10^{-7}$ is required.)

Phenocluster analysis in fly

As a further application of our variability metric, we next identify genes that modulate noise levels present in different morphological processes. To study these genes, we must first identify clusters of TCs (and by extension, genes) which are implicated in control of a particular morphological process (e.g. adhesion formation). By comparing the variability measure across all TCs within a given functionally-related cluster, we are able to identify genes that modulate morphological noise for a particular morphological process. For concreteness, we consider sets of TCs from the *Drosophila* screen thought to be involved in regulating a single morphological process. These so-called phenoclusters [8] are TCs for which the mean neural network Z scores forms a single cluster in (reduced) feature space; intuitively, the "average" morphology of cells in each of these TCs is similar. In [8], a total of 41 phenoclusters were identified, but we focus on the three largest phenoclusters, which correspond to adhesion disassembly/cortical tension, protrusion/adhesion formation, and lamellopodia formation. We calculated the variability p-scores for all TCs in each phenocluster and identify TCs with variability p-scores that are significantly elevated or depressed, relative to the other TCs in the same phenocluster.

To compare variability p-scores within phenoclusters rather than across the entire set of TCs, an alternate formulation for computing variability p-scores was executed. Note that variability v-scores were computed in the same way as before, i.e. as the (normalized) average of the squared distances of the points in the TC from their center of mass. In calculating the variability p-score for a TC of size n having variability v-score, v, we used the definition $\hat{p} = P(\widehat{Y_n} < v)$, where $\widehat{Y_n}$ denotes the distribution of variability v-scores drawn from $\hat{S}$, the set of single cells in the relevant phenocluster. This alternate definition accomplishes the goal of comparing variability among an underling set of single cells that resemble one another more closely in mean morphology than the full set, S.

To calculate variability p-scores for phenoclusters, we used bootstrapping with the same number of iterations (10^5 for *Drosophila*). The difference between the bootstrapping algorithm in this case was that randomly selected sets of n single cells were drawn from $\hat{S}$ instead of S. Note that the relative order of variability p-scores for TCs within a phenocluster is unchanged, whether we use $\hat{p}$ or p. However, using $\hat{p}$ gives a more accurate calculation of significance values for high/low phenocluster-specific variability.

References

1. Barkai N., Shilo B. Molecular Cell 28(5), 755-760 (2007)

2. Pedraza JM, van Oudenaarden A. Science. 2005 Mar 25;307(5717):1965-9.

3. Blake W.J., Kaern M., Cantor C.R., Collins J.J. Nature 422(6932), 633-7 (2003)

4. Raser J.M., O'Shea E.K. Science 304(5678), 1811-4 (2004)

5. Ozbudak E.M., Thattai M., Kurtser I., Grossman A.D., van Oudenaarden A. Nat Genet. 31(1), 69-73 (2002)

6. Carpenter A.E., Jones T.R., Lamprecht M.R., Clarke C., Kang I.H., Friman O., Guertin D.A., Chang J. H., Lindquist R., Moffat J., Golland P., Sabatini D. M. Genome Biol. 7(10), R100 (2006)

7. Ohya Y., Sese J., Yukawa M., Sano F., Nakatani Y., Saito T.L., et al. Proc Natl Acad Sci U S A. 102(52), 19015-20 (2005)

8. Bakal C., Aach J., Church G., Perrimon N. Science 316(5832), 1753-1756 (2007)

9. Moffat J., Grueneberg D.A., Yang X., Kim S.Y., Kloepfer A.M., Hinkle G., et al. Cell 124(6), 1283-98 (2006)

10. Glory E., Murphy R.F. Dev Cell 12(1), 7-16 (2007)

11. Perlman Z.E., Slack M.D., Feng Y., Mitchison T.J., Wu L.F., Altschuler S.J. Science 306(5699), 1194-8 (2004)

12. Neumann B., Held M., Liebel U., Erfle H., Rogers P., Pepperkok R. Nat Methods 3(5), 385-90 (2006)

13. Piano F., Schetter A.J., Morton D.G., Gunsalus K.C., Reinke V., Kim S.K., et al. Curr Biol. 12(22), 1959-64 (2002)

14. Gil J., Wu H., Wang B.Y. Microsc Res Tech 59(2), 109-18 (2002)

15. Levy SF, Siegal ML. Network Hubs Buffer Environmental Variation in Saccharomyces cerevisiae. PLoS Biol. 2008 Nov 4;6(11):e264.

16. Slack MD, Martinez ED, Wu LF, Altschuler SJ. Characterizing heterogeneous cellular responses to perturbations. Proc. Natl. Acad. Sci. U.S.A. 2008 Dec 9;105(49):19306-19311.

17. Spencer SL, Gaudet S, Albeck JG, Burke JM, Sorger PK. Non-genetic origins of cell-to-cell variability in TRAIL-induced apoptosis. Nature. 2009 May 21;459(7245):428-432.

18. The Gene Ontology Consortium. Nature Genet. 25: 25-29 (2000).

19. Versele M, Thorner J. J. Cell Biol. 2004 Mar 1;164(5):701-715.

20. Weiss EL, Bishop AC, Shokat KM, Drubin DG. Nat. Cell Biol. 2000 Oct ;2(10):677-685.

21. Bussey H. Science. 1996 Apr 12;272(5259):224.

22. Chant J, Stowers L. Cell. 1995 Apr 7;81(1):1-4.

23. Williams MJ, Habayeb MS, Hultmark D. J. Cell. Sci. 2007 Feb 1;120(Pt 3):502-511.

24. Hicks MS, O'Leary V, Wilkin M, Bee SE, Humphries MJ, Baron M. Dev. Genes Evol. 2001 May ;211(5):263-267.

25. Raj A, van Oudenaarden A. Cell. 2008 Oct 17;135(2):216-226.

26. Somers WG, Saint R. Dev. Cell. 2003 Jan ;4(1):29-39.

27. Dean SO, Rogers SL, Stuurman N, Vale RD, Spudich JA. Proc. Natl. Acad. Sci. U.S.A. 2005 Sep 20;102(38):13473-13478.

28. Zavortink M, Contreras N, Addy T, Bejsovec A, Saint R. J. Cell. Sci. 2005 Nov 15;118(Pt 22):5381-5392.

29. Longtine MS, Theesfeld CL, McMillan JN, Weaver E, Pringle JR, Lew DJ. Mol. Cell. Biol. 2000 Jun 1;20(11):4049-4061.

30. McMurray MA, Thorner J. Cell Cycle. 2009 Jan 15;8(2):195-203.

31. Versele M, Thorner J. Trends Cell Biol. 2005 Aug ;15(8):414-424.

32. Caviston JP, Longtine M, Pringle JR, Bi E. Mol. Biol. Cell. 2003 Oct 1;14(10):4051-4066.

33. Lee PR, Song S, Ro H, Park CJ, Lippincott J, Li R, Pringle JR, De Virgilio C, Longtine MS, Lee KS. Mol. Cell. Biol. 2002 Oct 1;22(19):6906-6920.

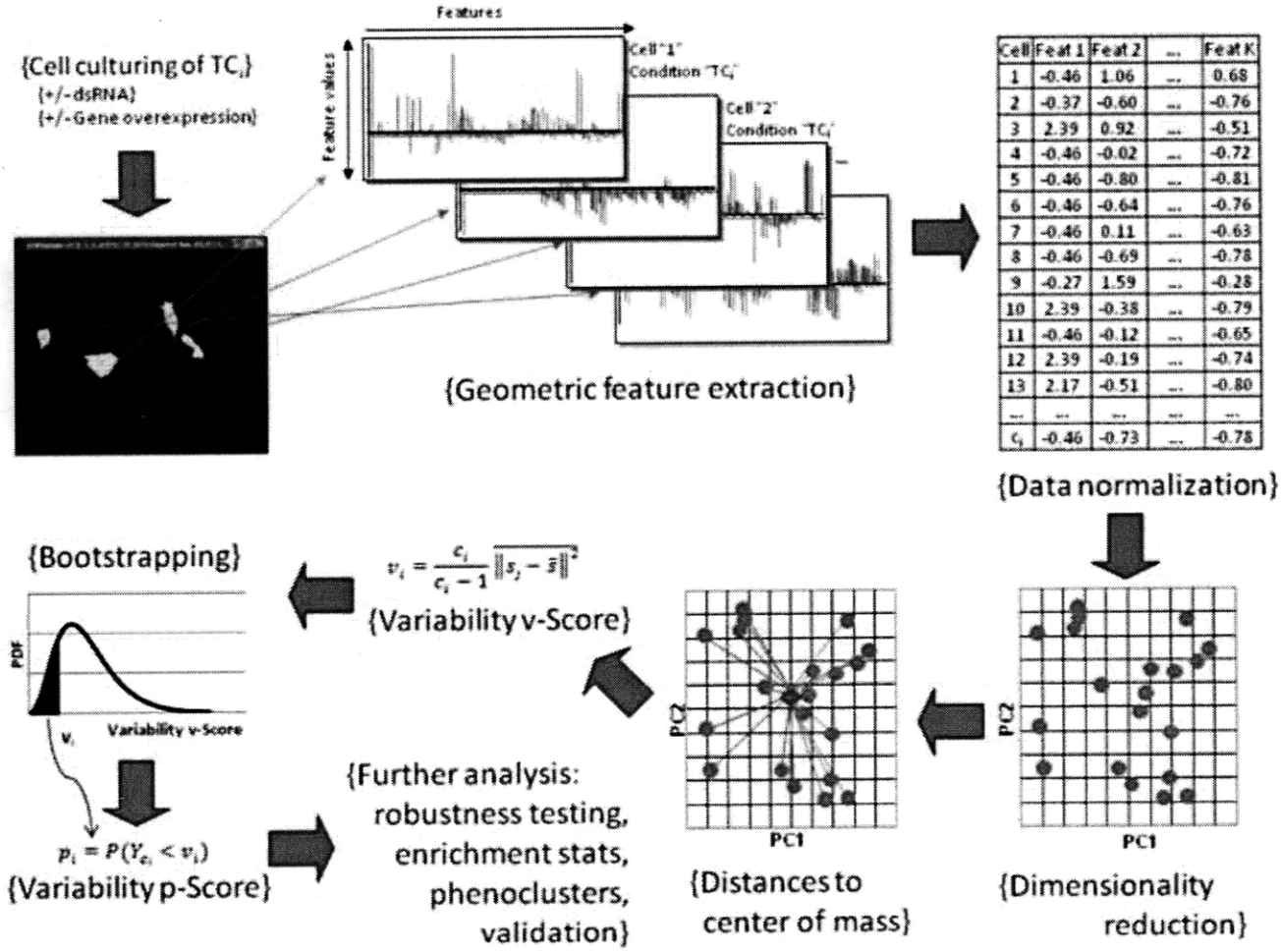
34. Bouquin N, Barral Y, Courbeyrette R, Blondel M, Snyder M, Mann C. Regulation of cytokinesis by the Elm1 protein kinase in Saccharomyces cerevisiae. J Cell Sci. 2000 Apr 15;113(8):1435-1445.

35. Longtine MS, Fares H, Pringle JR. J. Cell Biol. 1998 Nov 2;143(3):719-736.

36. Lew D, Reed S. J. Cell Biol. 1995 May 1;129(3):739-749.

37. Mahalanobis, P C. Proceedings of the National Institute of Sciences of India 1936, **2** (1): 49–55.

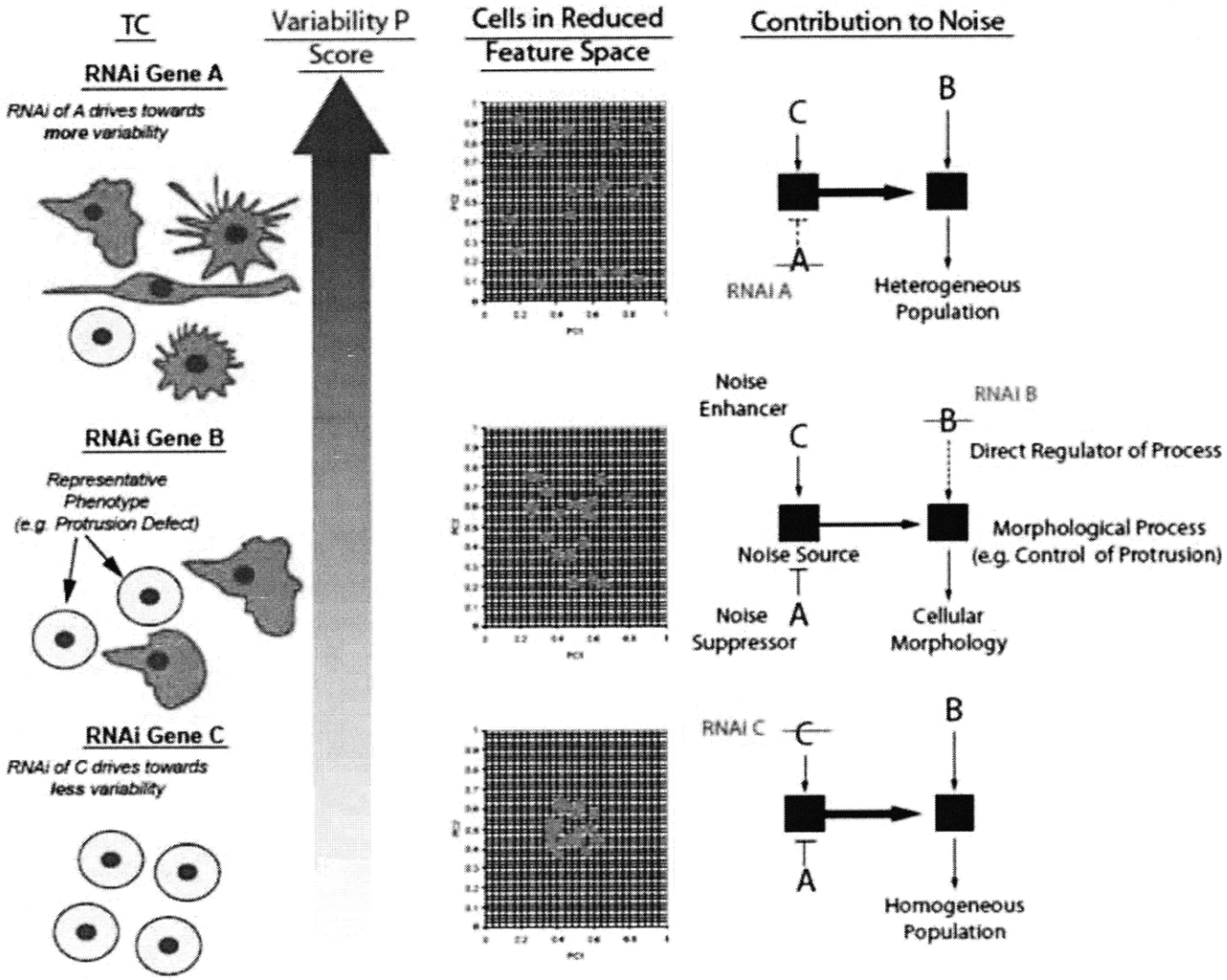
Figure 1: Workflow for computation of variability p-scores.



Cell culture was subjected to a variety of genetic perturbations (TCs), multiple single-cell images were acquired for each TC, and raw geometric features were extracted for each single cell in both [7] and [8] (upper left, upper middle). Using this raw feature data as our starting point, we first perform a normalization (top right) and dimensionality reduction of the raw data. The c_i single cells comprising each TC_i were represented as points in reduced feature space (bottom

right). The variability v-score for TC_i was computed according to the formula shown, which represents a normalized average of the squared distances of the points in TC_i from their center of mass (bottom middle). Subsequently, bootstrapping was performed to determine the distribution of variability v-scores for samples of size c_i drawn from the full set of single cells from all TCs; this distribution was used to calculate the variability p-score for TC_i (bottom left). Variability p-scores were calculated for all TCs and were subjected to further analysis. See text for additional details.

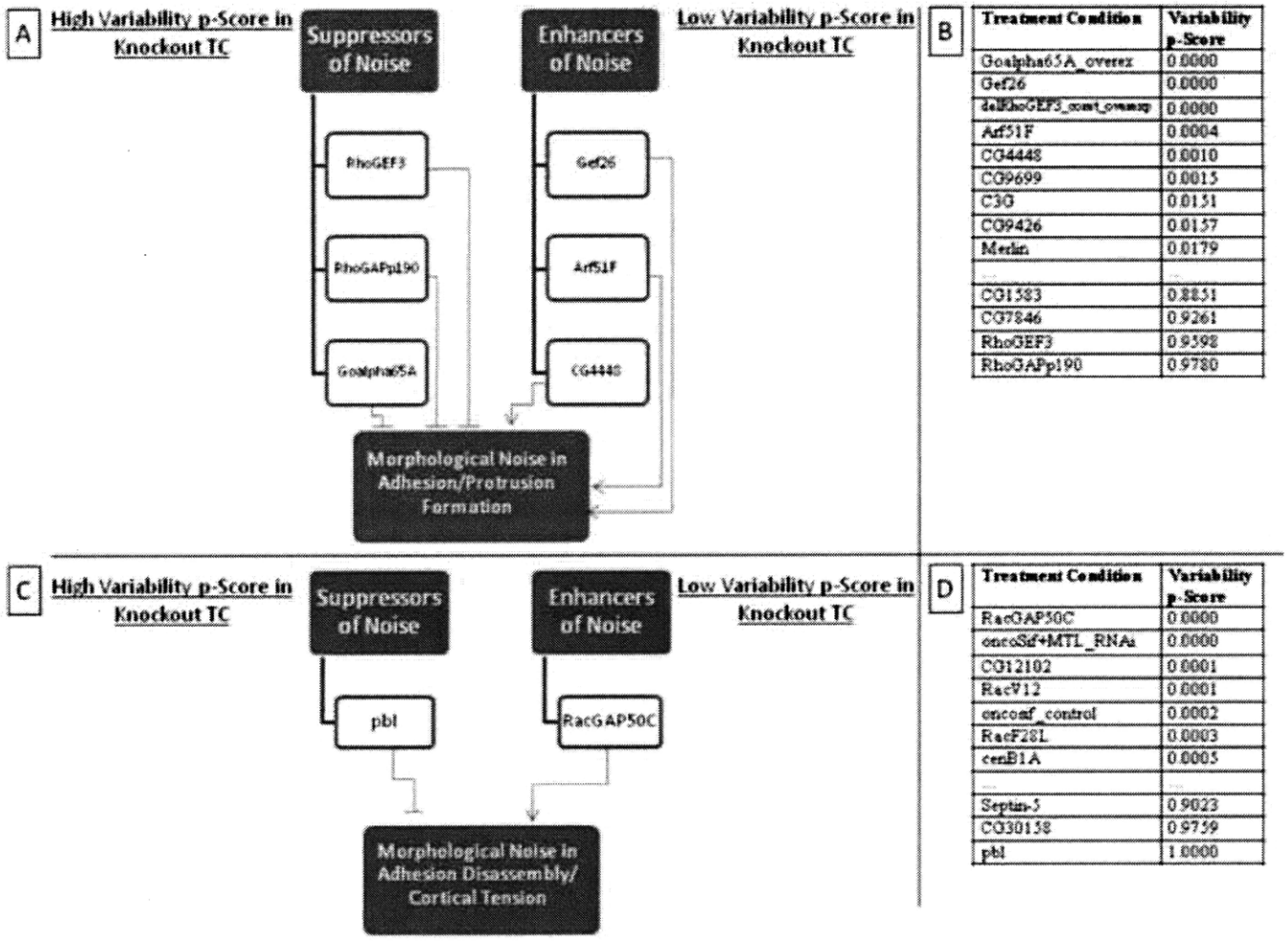
Figure 2: Using variability p-scores to quantify population variability and to determine the contribution of genes to morphological noise.



The first three columns summarize the relationship between population-level morphological variation, the variability p-score, and the representation of single cells as points in reduced feature space. In particular, as populations become more heterogeneous, the variability p-score increases, and the point set in reduced feature space increases in spread. The fourth column

introduces the concepts of suppression and enhancement of morphological noise. By systematically comparing variability p-scores for genes thought to be involved in control of a particular morphological process (genes A, B, and C), we identify genes which function either to increase or decrease noise in that process. Genes that increase (gene A) or decrease (gene C) morphological noise in the population when inhibited by RNAi are designated as noise suppressors or enhancers, respectively, of a particular morphological process. These genes are identified by variability p-scores which are either high or low, to statistical significance. Genes that when inhibited by RNAi do not result in statistically significant variability p-scores are considered direct regulators of a cellular process (gene B).

Figure 3: Variability analysis in *Drosophila* phenoclusters.



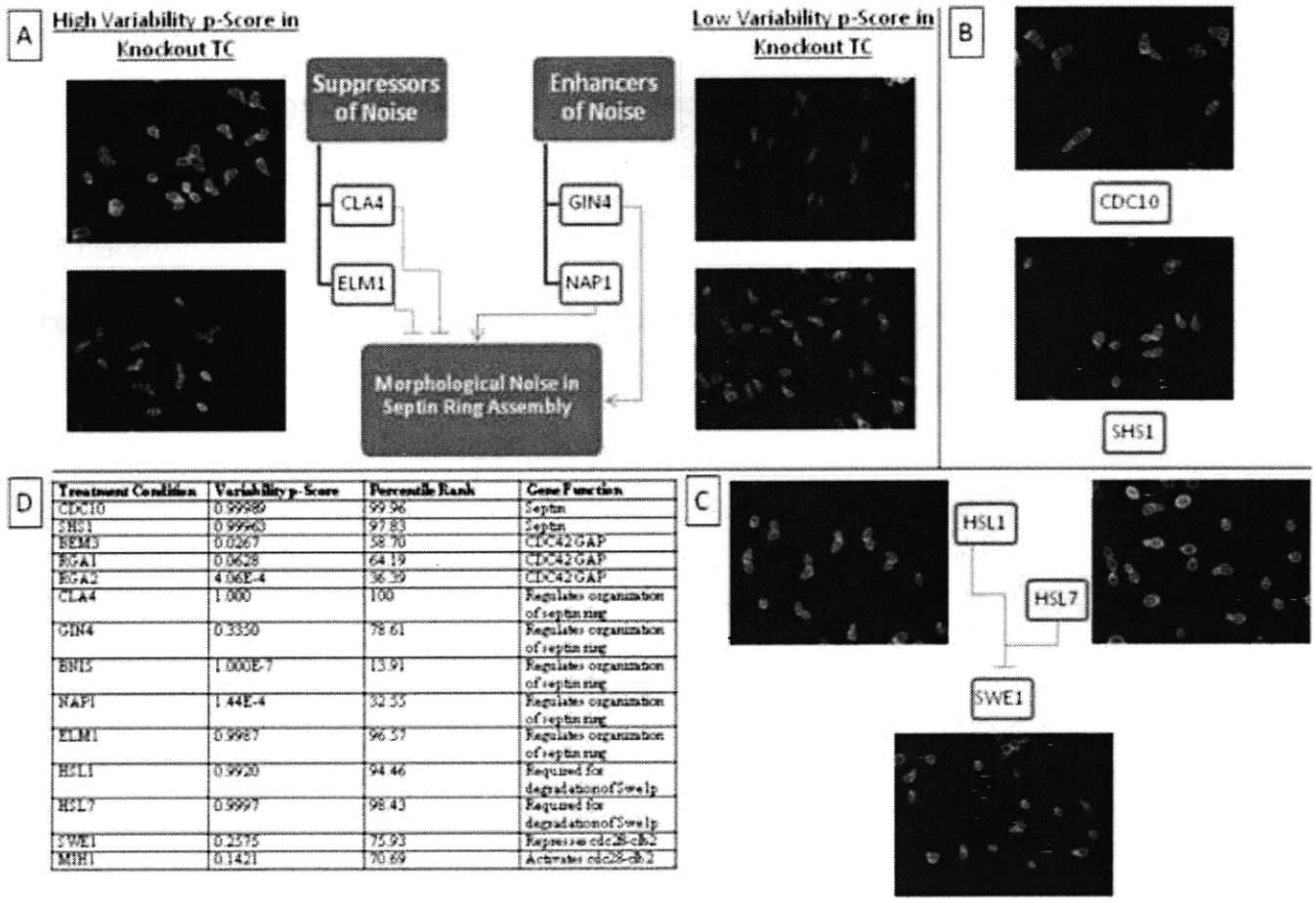
Treatment Condition	Variability p-Score
Gosalpha65A_overexp	0.0000
Gef26	0.0000
delRhoGEF3_contr_overexp	0.0000
Arf51F	0.0004
CG4448	0.0010
CG9699	0.0015
C3G	0.0151
CG9426	0.0157
Merlin	0.0179
...	...
CG1583	0.8851
CG7846	0.9261
RhoGEF3	0.9398
RhoGAPp190	0.9780

Treatment Condition	Variability p-Score
RacGAP50C	0.0000
encoSif+MTL_RNAi	0.0000
CG12102	0.0001
Rac V12	0.0001
encomf_control	0.0002
RacF28L	0.0003
cenB1A	0.0005
...	...
Septin5	0.9023
CG30138	0.9759
pbl	1.0000

Selected variability results for TCs in the *Drosophila* phenoclusters for protrusion/adhesion formation (panels A, B) and for adhesion disassembly/cortical tension (panels C, D). (A) Suppressors and enhancers of noise for protrusion/adhesion formation. TCs defined by RNAi of *RhoGEF3* and *RhoGAPp190* each resulted in elevated variability p-scores; these scores were not significant following Bonferroni correction, but are noteworthy because these were the highest variability p-scores among all TCs in the protrusion/adhesion phenocluster. Likewise, RNAi of

Gef26, *Arf51F*, and *CG4448* and overexpression of *Goalpha65A* resulted in significantly reduced variability p-scores. (B) Variability p-scores for high- and low-scoring TCs from the protrusion/adhesion formation phenocluster. (C) Suppressors and enhancers of noise for adhesion disassembly/cortical tension. Knockout of *pbl* resulted in significantly elevated morphological variability, while knockout of *RacGAP50C* resulted in significantly lowered morphological variability. (D) Variability p-scores for high- and low-scoring TCs from the adhesion disassembly/cortical tension phenocluster. See text and **Supplementary Figs. 3-6** for additional details.

Figure 4: Variability results for TCs for yeast genes involved in regulating septin ring formation.



Treatment Condition	Variability p-Score	Percentile Rank	Gene Function
CDC10	0.9989	99.96	Septin
SHS1	0.9963	97.83	Septin
BEM3	0.0267	58.70	CDC42GAP
RGA1	0.0628	64.19	CDC42GAP
RGA2	4.06E-4	36.39	CDC42GAP
CLA4	1.000	100	Regulates organization of septin ring
GIN4	0.3350	78.61	Regulates organization of septin ring
BNI5	1.000E-7	13.91	Regulates organization of septin ring
NAP1	1.44E-4	32.55	Regulates organization of septin ring
ELM1	0.9987	96.57	Regulates organization of septin ring
HSL1	0.9920	94.46	Required for degradation of Swe1p
HSL7	0.9997	98.43	Required for degradation of Swe1p
SWE1	0.2275	75.93	Represses cdc28-clb2
MIH1	0.1421	70.69	Activates cdc28-clb2

(A) The genes CLA4, ELM1, GIN4, and NAP1 were known to regulate septin ring assembly. Variability p-scores for CLA4 and ELM1 were among the highest (CLA4 was the single highest) among the full set of 4787 yeast TCs in the genetic screen, while the variability p-scores for NAP1 and GIN4 were not significantly elevated. See text for additional details. (B) TCs defined by knockout of two of the septins (CDC10, SHS1) resulted in single cell populations with

significantly elevated morphological variability, while knockout of either of the three other septins is lethal. (C) Knockout of HSL1 or HSL7 results in significantly elevated population variability, while knockout of SWE1 yields a variability p-score that is not significantly elevated. This is consistent with the known network architecture of Swe1p regulation (see text for further details). (D) Table of variability p-scores (second column) for TCs (first column) defined by knockout of genes involved in septin ring regulation. The third column reports the percentile rank for the variability p-score, relative to the entire set of 4787 TCs, and the fourth column summarizes gene function. Genes implicated in septin ring regulation mentioned in the text but not included in this table were not included in the genetic screen. See text and **Supplementary Figs. 7-9** for further discussion.

Table 1: TCs with significantly high morphological variability

Table 1A: Yeast TCs displaying high morphological variability

GO Term	P value for enrichment	Yeast genes
Chromosome organization	2.42e-07	RAD52 CYC8 DEF1 CTF8 BUD32 SPT10 CTF4 NPL6 SPC72 RTT109 RAD54 SCP160 HTL1 EST1 CIK1 RAD50 DCC1
Response to DNA damage stimulus	3.18e-07	RAD27 RAD52 RNR1 DEF1 SAC3 CTF8 SPT10 CTF4 NPL6 RTT109 RAD54 HTL1 HOF1 RAD50 DCC1
Cellular component organization	2.39e-03	BEM1 RPB4 RAD52 CYC8 EDC3 DEF1 SAC3 SHE4 CTF8 BUD32 IRC25 CDC10 SPT10 CTF4 NPL6 SPC72 RTT109 RAD54 SCP160 HTL1 EST1 CHC1 CLA4 CIK1 BEM2 RAD50 DCC1
Nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	2.44e-03	CTK2 RPB4 RNR4 RAD27 RAD52 CYC8 RPA12 SWI4 EDC3 ADH1 RNR1 DEF1 SAC3 SQS1 SHE4 BUD32 IRC25 SPT10 CTF4 NPL6 RTT109 RAD54 SCP160 HTL1 EST1 ANP1 RAD50
Anatomical structure homeostasis / telomere maintenance / telomere organization	4.23e-03	RAD52 DEF1 BUD32 RAD54 EST1 RAD50
Organelle organization	5.71e-03	RAD52 CYC8 DEF1 SAC3 SHE4 CTF8 BUD32 SPT10 CTF4 NPL6 SPC72 RTT109 RAD54 SCP160 HTL1 EST1 CLA4 CIK1 BEM2 RAD50 DCC1
Cell cycle	7.79e-03	SAC3 CTF8 CTF4 SPC72 CLA4 CIK1 DCC1
Mitosis / nuclear division /	9.52e-03	SAC3 CTF8 CTF4 SPC72

phase of mitotic cell cycle		CLA4 CIK1 DCC1

Gene Ontology enrichment statistics for the yeast genes scoring in the top 1% on our variability metric. This set of 47 yeast TCs was significantly enriched for 12 GO categories. These categories are shown above in the leftmost column, condensed into 8 categories (due to redundancy within GO, as shown). The P values, as shown in the middle column, already incorporate Bonferroni correction. Finally, the rightmost column lists the yeast genes from the relevant GO category. See **Materials and Methods** for computational details.

Table 1B: Fly TCs displaying high morphological variability

Treatment Condition	Variability p-Score
Septin-5	0.9580
CG8557	0.9722
CG7324	0.9827
Cdc42	0.9905
CG30158	0.9934
Pbl	1.0000

TCs from the *Drosophila* screen with population variability that is increased relative to random. The top ranking TCs are shown ($p > 1 - .05$), though only the highest-scoring TC, *pbl*, is significant after Bonferroni correction ($p > 1 - .05/273$). In the left column is the TC name. In the right column, the variability p-score for each TC is shown. *Pbl* encodes a protein required for cells to undergo cytokinesis. Thus, cells lacking *pbl* become bi/multi-nucleated and display an enormous amount of morphological variability. *Cdc42* encodes a protein that normally regulates the formation of highly dynamic protrusive events, such as the formation of filopodia, at the leading edges of cells. Our finding that knocked-down *Cdc42* induces increased variability suggests that protrusive events become unregulated in the absence of its protein product.

Table 2: TCs with significantly low morphological variability

Table 2A: Yeast TCs displaying low morphological variability

GO Term	P value for enrichment	Yeast genes
Mitochondrial translation	3.32e-03	ISM1 MSW1 MRPS12 MRPL23 NAM2 RSM25 RSM18 MRPL11 IFM1 RTC6 RSM19 MRPL32 MRPS8 MRPL10 MRPL16 MRPL7 MRP10 SWS2 MRPL13 RSM27 MSE1 GRS1 MSR1 PET112

GO enrichment statistics for the lowest-scoring yeast TCs. A total of 491 genes scored at $p < 10^{-8}$ on our variability metric (meaning that none of the bootstrapped samples scored lower on the variability metric), which corresponds to Bonferroni-corrected $p < 5 \cdot 10^{-5}$. This set is enriched for genes involved in mitochondrial translation, and these genes are shown in the rightmost column. Computation details for the GO enrichment calculations are included in **Materials and Methods**.

Table 2B: Fly TCs displaying low morphological variability

Treatment Condition	Variability p-Score	Phenocluster
GEF64C	< 1E-5	Lamellipodia formation
Graf	< 1E-5	Lamellipodia formation
RhoF30L	< 1E-5	Lamellipodia formation
twinstar	< 1E-5	Lamellipodia formation
Trio	1E-4	Lamellipodia formation
Goalpha65A	< 1E-5	Protrusion/Adhesion formation
Merlin	< 1E-5	Protrusion/Adhesion formation
C3G	< 1E-5	Protrusion/Adhesion formation
CG9426	< 1E-5	Protrusion/Adhesion formation
CG9699	< 1E-5	Protrusion/Adhesion formation
CG4448	< 1E-5	Protrusion/Adhesion formation
CG7578	< 1E-5	Protrusion/Adhesion formation
Rab9	< 1E-5	Protrusion/Adhesion formation
armadillo	< 1E-5	Protrusion/Adhesion formation
Gef26	< 1E-5	Protrusion/Adhesion formation
Arf51F	< 1E-5	Protrusion/Adhesion formation
RhoGEF3	< 1E-5	Protrusion/Adhesion formation
delRhoGEF3_const_overex	< 1E-5	Protrusion/Adhesion formation
Ankyrin	1E-4	Protrusion/Adhesion formation
apc	< 1E-5	MT capture
apc2	1E-4	MT capture
RacGAP50C	< 1E-5	Adhesion disassembly/Cortical tension
CG12102	< 1E-5	Adhesion disassembly/Cortical tension
oncoSif+MTL_RNAi	< 1E-5	Adhesion disassembly/Cortical tension
dLis1_overex	< 1E-5	GFP/Wildtype
Rho1	< 1E-5	Rho1
oncoSif+Rho1_RNAi	< 1E-5	N/A

TCs from the *Drosophila* screen with population variability that is reduced relative to random ($p < .05/273$, using Bonferroni correction). In the leftmost column is the TC name. In the middle column, the variability p-score for each TC is shown; these were obtained by using bootstrapping with 10^4 iterations. In the rightmost column is recorded the phenocluster from [8] that each TC belongs to. Out of the 28 TCs with significantly reduced population variability, 19

are involved in either lamellipodia formation or protrusion/adhesion formation, while 63 of the original 273 TCs belong to the union of these two categories. The probability of this overlap occurring by chance is $< 9 \cdot 10^{-9}$, as given by the hypergeometric cdf, indicating significant enrichment. (This is true even after correcting for multiple hypotheses, as there 820 pairwise combinations of the 41 phenoclusters and, to be extremely conservative, 273 choices for where to draw a cutoff for inclusion in a group of lowest-scoring TCs, meaning that $p < \frac{.05}{820 \cdot 273} = 2 \cdot 10^{-7}$ is required.)

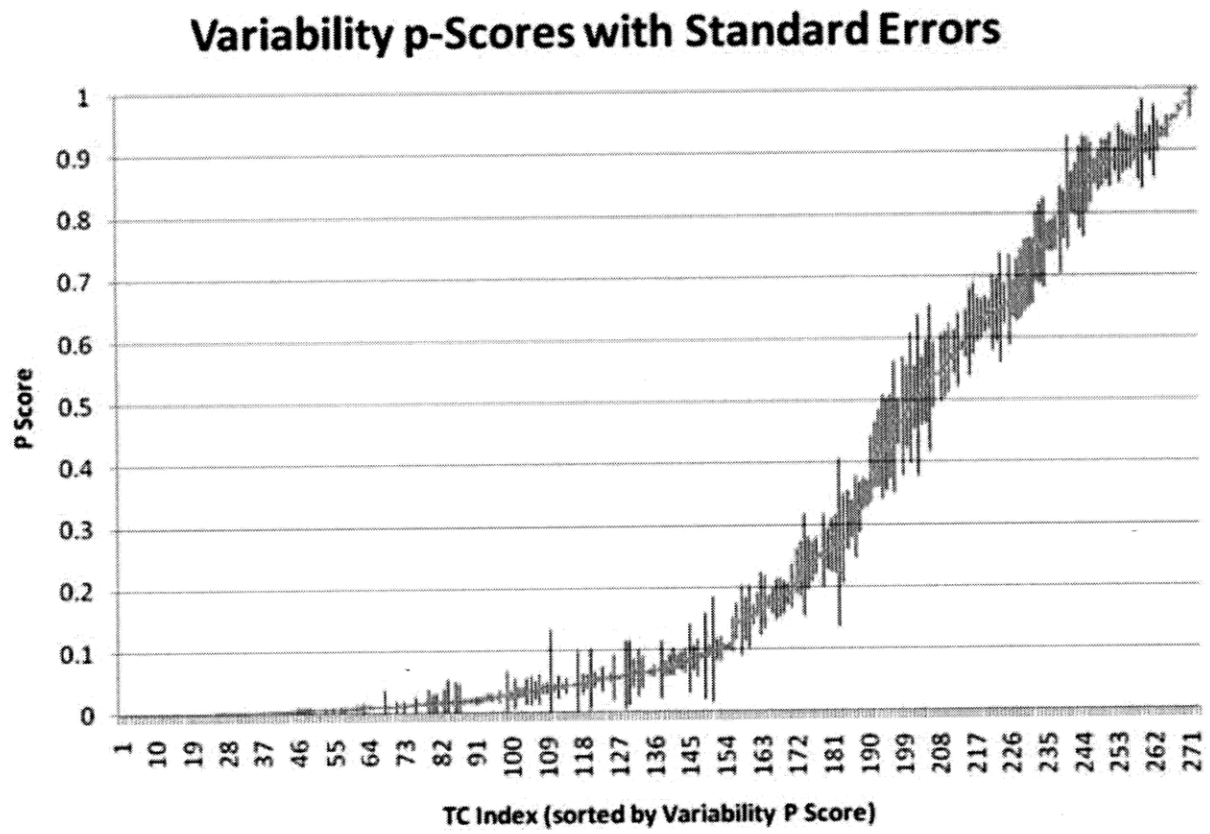
Genetic Tuning of Morphological Variability in Cellular Processes

Oaz Nir, Chris Bakal, Norbert Perrimon & Bonnie Berger

Supplementary figures and text:

Supplementary Figure 1

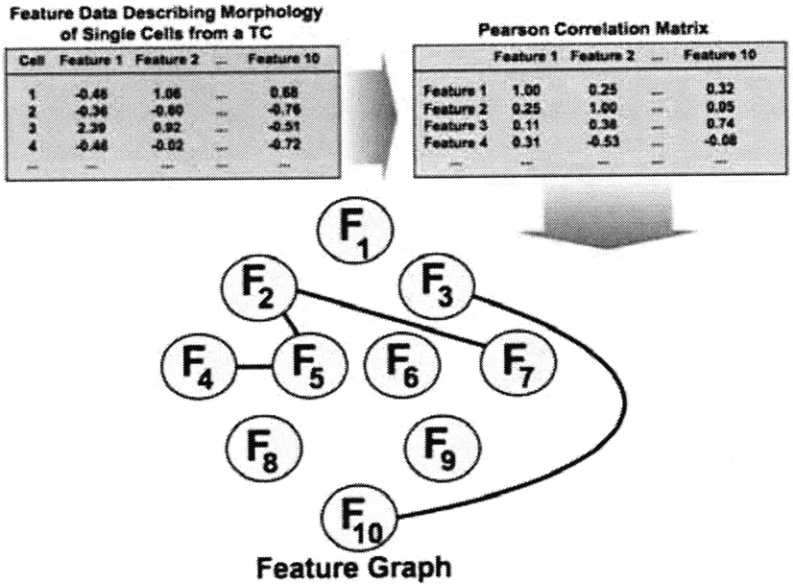
Standard errors for variability p-scores in *Drosophila* from jackknifing



Supplementary Fig. 1. Jackknife statistics for all TCs in the *Drosophila* screen were computed, which allowed us to calculate standard errors for each TC's variability v-score and variability p-score. TCs were sorted according to variability p-score, and error bars are as shown.

Supplementary Figure 2

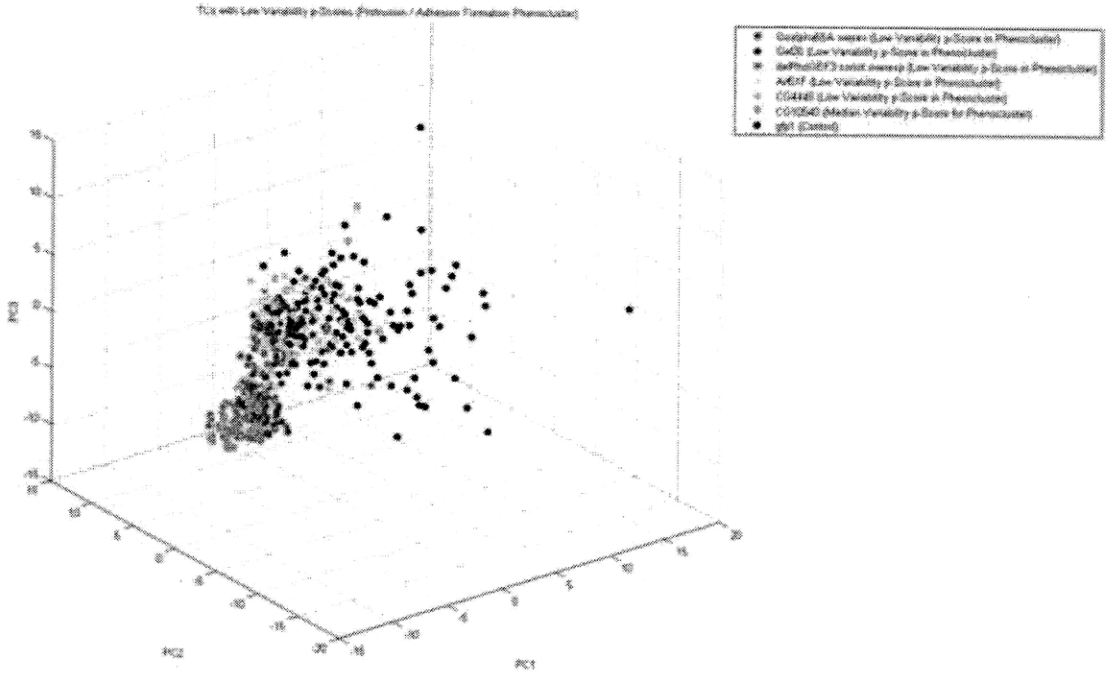
Alternate approach to variability measurement using pairwise feature correlations (Feature Graphs)



Supplementary Fig. 2. Steps in Feature Graph Construction for a Single TC. Pairwise Pearson correlations are computed for each pair of features, taken across all cells in the TC. If a correlation between two features is sufficiently large in magnitude (relative to a determined threshold) then the vertices corresponding to these features are connected by an edge in the FG. In this example, the dimensionality of reduced feature space is $k = 10$.

Supplementary Figure 3

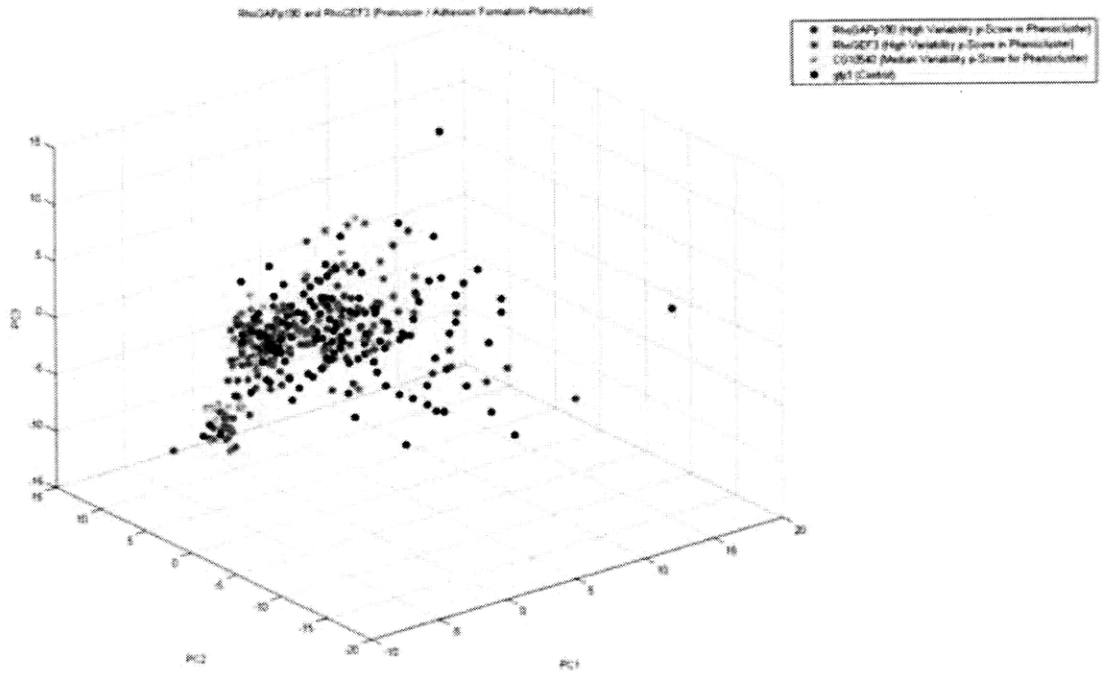
TCs with significantly low variability p-scores for protrusion/adhesion formation in fly



Supplementary Fig. 3. TCs with statistically significant, low variability p-scores for the protrusion adhesion phenocluster in PC-based coordinates. The plot shows qualitatively what the variability metrics make mathematically rigorous, namely the extent of spread in the lowest scoring TCs for this phenocluster as well as two reference TCs (a control TC; and CG10540, which had the median variability p-score for the phenocluster). Note that the point set for delRhoGEF3 displays qualitatively less spread than all the other low scoring TCs.

Supplementary Figure 4

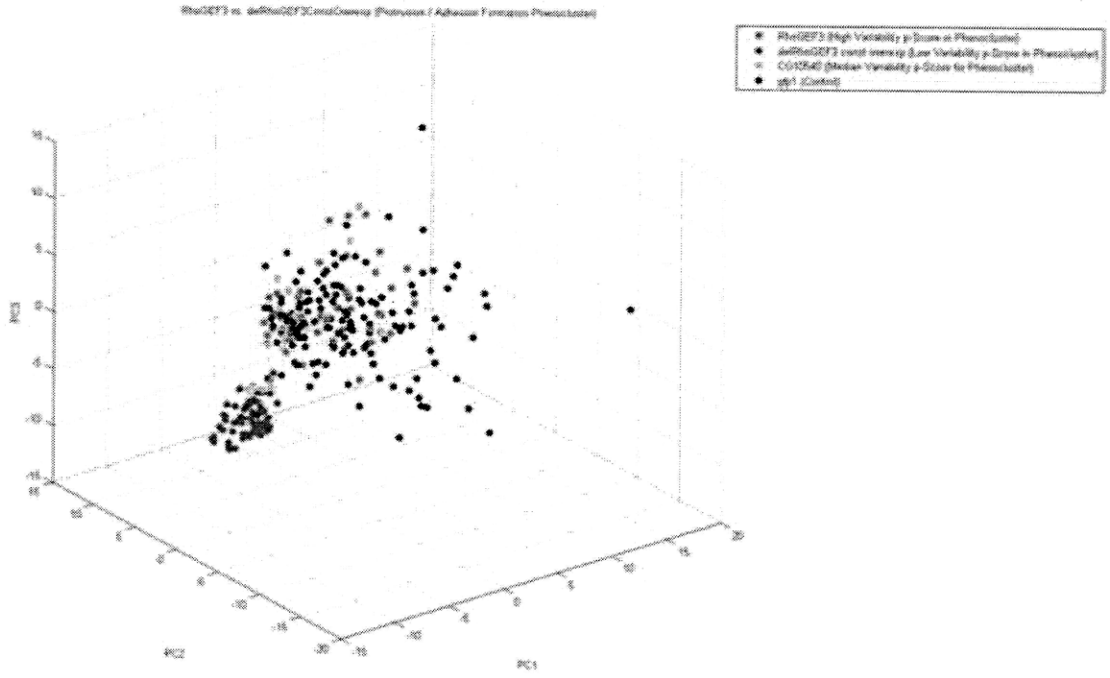
TCs with significantly high variability p-scores for protrusion/adhesion formation in fly



Supplementary Fig. 4. TCs with high variability p-scores for the protrusion/adhesion formation phenocluster. Point sets in PC-based coordinates are shown for RhoGAPp190 and RhoGEF3 (the two highest scorers in this phenocluster) as well as two references (a control TC; and CG10540, which had the median variability p-score for this phenocluster). Note that the point sets for RhoGAPp190 and RhoGEF3 are of comparable spread to the control TC, which reflects the fact that the TCs in this phenocluster have low morphological variability compared to the full set of TCs (see main text for further discussion of this point). See **Supplementary Fig. 6** as a point of comparison in this regard.

Supplementary Figure 5

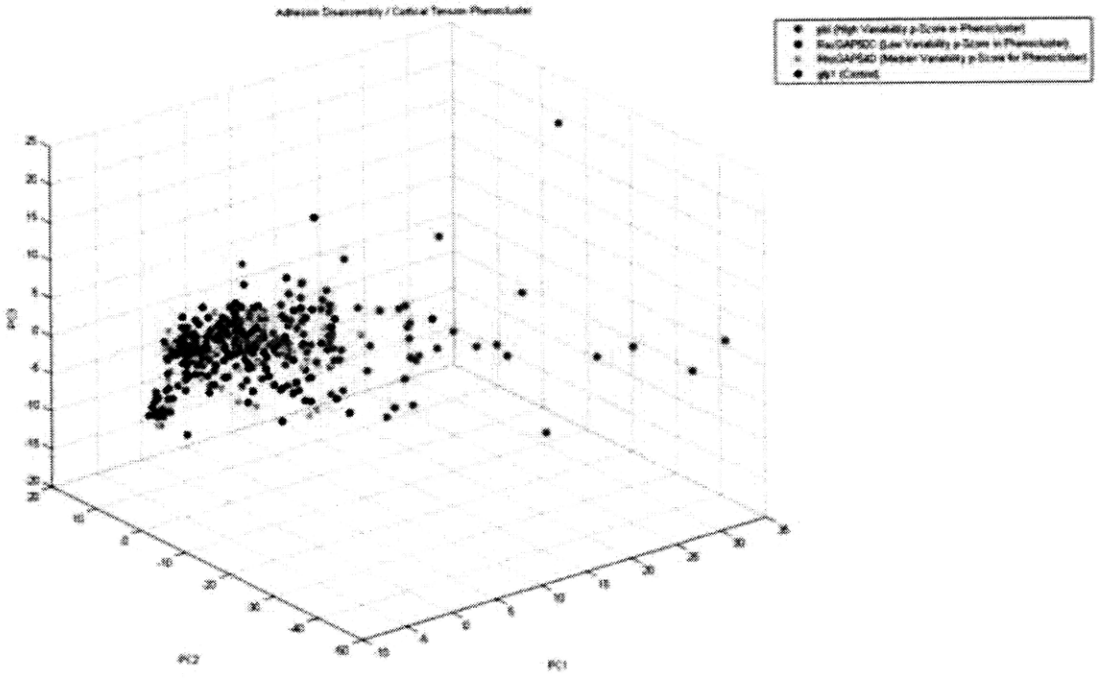
RhoGEF3 and delRhoGEF3_const_overexp for protrusion/adhesion formation in fly



Supplementary Fig. 5. RhoGEF3 and delRhoGEF3_const_overexp for the protrusion/adhesion formation phenocluster. Point sets in PC-based coordinates are shown for these two TCs as well as two references (a control TC; and CG10540, which had the median variability p-score for this phenocluster). The RhoGEF3 knockout population displays high morphological variability, while the overexpression population displays low variability, supporting the role of RhoGEF3 as a suppressor of morphological noise.

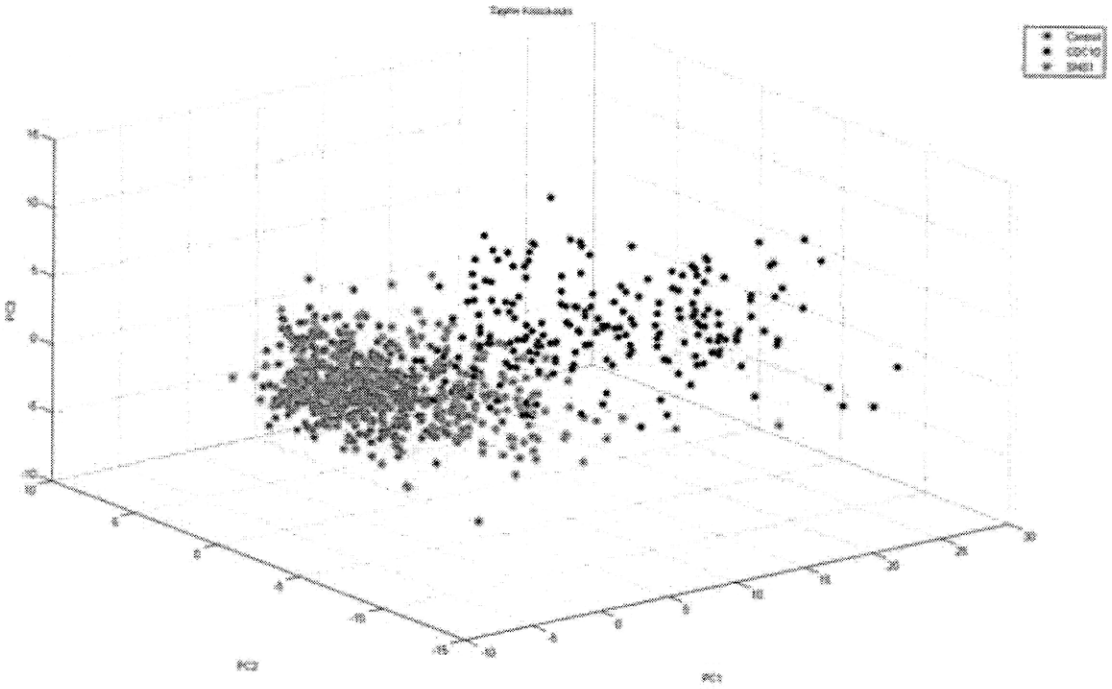
Supplementary Figure 6

TCs with significant variability p scores for adhesion disassembly/cortical tension in fly



Supplementary Fig. 6. TCs with statistically significant variability p scores for the adhesion disassembly/cortical tension phenocluster. Point sets in PC-based coordinates are shown for RacGAP50C and pbl as well as two references (a control TC; and RhoGAP54D, which had the median variability p-score for this phenocluster). The most extreme points in the pbl set correspond to several different RNAi constructs. As a point of comparison, see **Supplementary Fig. 4**, in which the highest scoring TCs for the protrusion/adhesion formation phenocluster had less variability than the highest scoring TCs (e.g. pbl) for the adhesion disassembly/cortical tension phenocluster shown here.

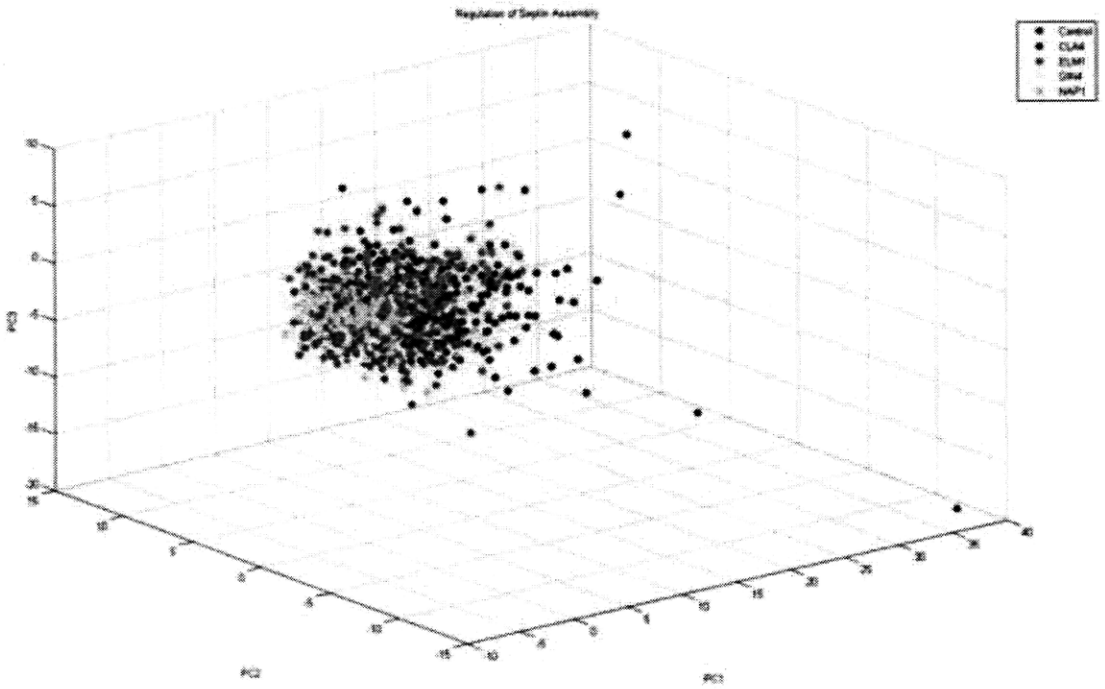
Supplementary Figure 7
Septin knockout TCs in yeast



Supplementary Fig. 7. Septin knockout TCs in yeast. Point sets in PC-based coordinates are shown for CDC10 and SHS1 knockout TCs as well as a control TC for reference. Both septin knockout TCs display greater morphological variability – see main text and **Fig. 4** for further details.

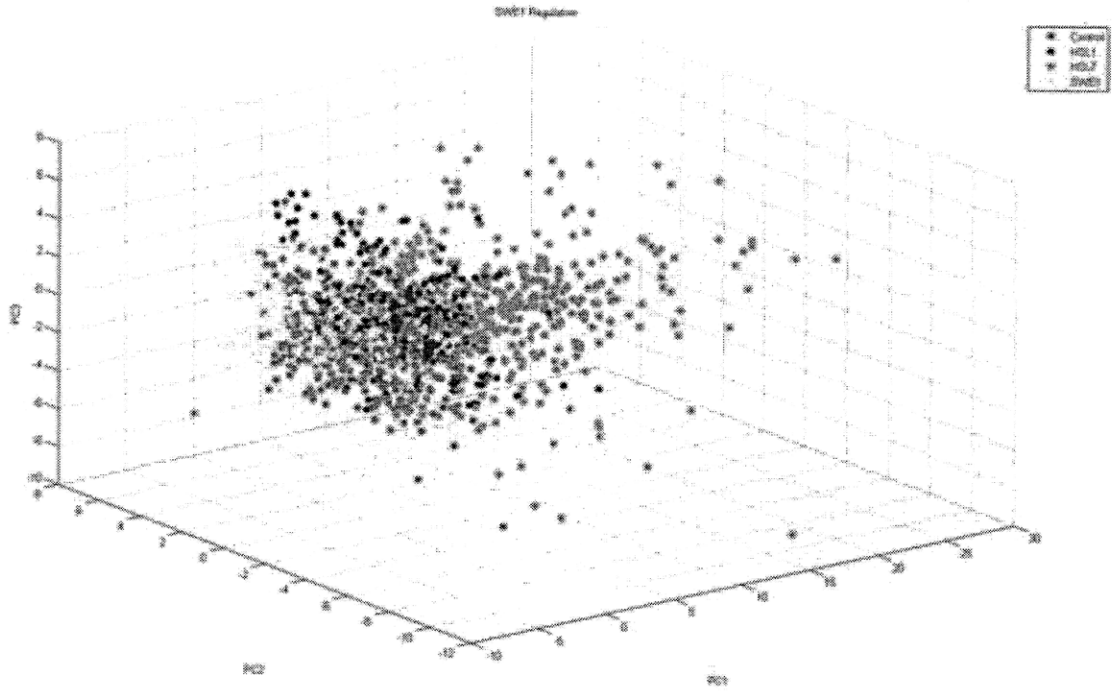
Supplementary Figure 8

Regulation of septin assembly in yeast



Supplementary Fig. 8. Regulation of septin assembly in yeast. Point sets in PC-based coordinates are shown for CLA4, ELM1, GIN4, and NAP1 knockout TCs as well as a control TC for reference. The CLA4 and ELM1 knockout TCs have high morphological variability, whereas GIN4 and NAP1 display low morphological variability, supporting the role of these genes as suppressors and enhancers of morphological noise, respectively. See main text and **Fig. 4** for further details.

Supplementary Figure 9
SWE1 regulation in yeast



Supplementary Fig. 9. SWE1 regulation in yeast. Point sets in PC-based coordinates are shown for HSL1, HSL7, and SWE1 knockout TCs as well as a control TC for reference. The HSL1 and HSL7 knockout TCs have high population variability whereas the SWE1 knockout has low variability, reflecting the architecture of septin regulation. See main text and **Fig. 4** for further details.

Supplementary Table 1

List of treatment conditions from the *Drosophila* screen

Treatment Condition	Function
Rac1	Rho GTPase
GEF64C overexpression	RhoGEf
Rac1Rac2MTL	Rho GTPase
hAuroraB const. active overexpression	
CG32627	
G protein beta 76C	G-gamma
Microtubule-associated protein 205	
CG8707	Rag GTPase
RhoGAP71E	RhoGAP
sec23	
GXIVsPLA2	
CG7940	
armadillo	
Centrosomal protein 190kD	
l(1)dd4	
Sop2	
lightoid	Rab GTPase
staufen	
slingshot	
Ankyrin	
Arf51F	Arf GTPase
Merlin	
rho-like	Rho GTPase
Gef26	RapGEF
cib	
C3G	RapGEF
RhoGAPp190	RhoGAP
CG8801	
CG7578	ArfGEF
CG1583	
CG9699	Septin GTPase
CG7846	
CG4267	
Rab30	Rab GTPase
CG5160	
CG5337	TBC GTPase
CG9426	
Rab9	Rab GTPase
CG9248	

CG4853	RasGEF
CG10540	
RhoGEF3	RhoGEF
CG33232	
CG6838	
CG4448	
SCAR	Rho effector
RapGAP1	RapGAP
G protein alpha-i 65A overexpression	G-beta
del-N-RhoGEF3 overexpression	RhoGEF
yurt	
Rab-protein 3	Rab GTPase
Neurofibromin 1	RasGAP
CG6017	
CG1193	
kelch	
RanGAP	RanGAP
alpha-Catenin	
twinstar	
cnn	
Septin-2	Septin GTPase
Trio	RhoGEF
Grip75	
Rab3-GAP	Rab GTPase
capt	
CSN1a	
CG3009	
Marf	
Graf	RhoGAP
Rab35	Rab GTPase
Arc-p20	
gartenzwerg	ArfGEF
CG15611	RhoGEF
Mapmodulin	
CG15097	
GEF64C	RhoGEF
RhoGAP68F	RhoGAP
Rab26	Rab GTPase
CG32030	
RhoF30L overexpression	Rho GTPase
Mtl	Rho GTPase
Vav	RhoGEF
CG5745	TBA GTPase
Cdep	RhoGEF
CG7324	RabGAP

del-N-SIF overexpression	RhoGEF
canoe	Rap effector
ran-like	Ran GTPase
Microtubule-associated protein 60	
miranda	
jitterbug	
p16-ARC	
del-N-SIF_RhoGEF3dsRNA	
concertina	G-alpha
Actinin	
G protein alpha-i 65A	G-alpha
G protein beta 13F	G-beta
Mp20	
pbl	RhoGEF
gamma-tubulin at 23C	
G protein alpha 49B	G-alpha
gamma tubulin 37C	
Gelsolin	
ADP ribosylation factor 79F	Arf GTPase
Moesin	
sanpodo	
Actin-related protein 66B	
Rab-protein 2	Rab GTPase
Rab5	Rab GTPase
Rab-RP4	Rab GTPase
locomotion defects	
Phospholipase A2 activator protein	
CG14782	
Brahma associated protein 55kD	
Rap21	
Crag	Rap GTPase
Septin-5	
Grip84	Septin GTPase
mini spindles	
falten	
centaurin gamma 1A	ArfGAP
alpha-catenin related	
Patsas	
lava lamp	
pod1	
CG12102	
CG11063	
RhoGAP15B	RhoGAP
RhoGAP19D	RhoGAP
CG13692	ArfGAP

CG9135/RCC-1	RanGEF
Menin 1	
Arc-p34	
CG9243	
Grp1	ArfGEF
CG12736	
CG16728	ArfGAP
Dystrobrevin-like	
RacGAP50C	RhoGAP
CG8479	Dynamin GTPase
CG5522	Ral GEF
CG15609	
RhoGAP54D	RhoGAP
EfSec	
CG33275	RhoGEF
CG10971	
CG10724	
CG7365	
Sar1	Sar GTPase
cenB1A	ArfGAP
RhoGAP100F	RhoGAP
paxillin	
CG18858	
CG30158	
CG30440	RhoGEF
CG30456	RhoGEF
CG31683	
MICAL	
RacF28L overexpression	Rho GTPase
RacV12 overexpression	Rho GTPase
SIF full-length overexpression	RhoGEF
del-N-SIF + GFP dsRNA	
del-N-SIF_CG3799dsRNA	
del-N-SIF_Rab5dsRNA	
del-N-SIF_RhoGAP16FdsRNA	
del-N-SIF_RhoGAP54D dsRNA	
del-N-SIF+MTL_RNAi	
homolog of RecQ	
SIF	RhoGEF
G protein gamma 30A	G-gamma
CG5022	
RhoBTB	Rho GTPase
formin 3	G-alpha
enabled	Rho effector
CG14034	

CG8243	ArfGAP
cappuccino	Rho effector
G protein s-alpha 60A	G-gamma
shibire	Dynamin GTPase
capping protein beta	
CG7420	RanGEF
mbc	RhoGEF
CG7787	RabGEF
Rheb	Rheb GTPase
del-N-SIF+Rho1_RNAi	
RabX2	Rab GTPase
CG14045	
dia	
apc	
apc2	
Fimbrin	G-alpha
CG32138	Rho effector
Muscle-specific protein 300	
chrowded	
CG8557	RhoGEF
CG10188	RhoGEF
Elongation factor 1?48D	
no receptor potential A	
Cdc42	Rho GTPase
rtGEF	RhoGEF
par1	
CG3799	RhoGEF
CG11490	TBC GTPase
Pld	
CG11968	Rag GTPase
CG12241	TBC GTPase
RhoGAP92B	RhoGAP
CG30115	RhoGEF
Moody beta overexpression	GPCR
Cdc42Y32A overexpression	Rho GTPase
CG3799 overexpression	RhoGEF
RhoV14	Rho GTPase
del-N-SIF+Rac1_RNAi	
dLis1 overexpression	
dPar1 overexpression	
dStrad overexpression	
gfp1	
gfp2	
Nrg overexpression	
TumL overexpression	

visceral mesodermal armadillo-repeats	
G protein o-alpha 47A	G-gamma
Sos	RhoGEF
alpha-Spectrin	
Rho1	Rho GTPase
RhoGEF2	RhoGEF
RhoGAP1A	RhoGAP
RhoGAP5A	RhoGAP
RhoGAP16F	RhoGAP
RhoGAP18B	RhoGAP
RhoGEF4	Rho GTPase
CG7323	RhoGEF
RhoGAP93B/Vilse	RhoGAP
RhoGAP102A	RhoGAP
CG30372	ArfGAP
RabX4	Rab GTPase
CdGAPr	RhoGAP
del-N-SIF_Arcp34dsRNA	
del-N-SIF_EnadsRNA	
dMemo overexpression	
Rab-protein 7	Rab GTPase
peanut	
CG7197	Arf GTPase
Bj1 protein	
CLIP-190	
CG14507	
Rab-protein 6	Rab GTPase
CG8397	
G protein gamma 1+A27	
abnormal spindle	

Supplementary Table 1. Treatment conditions (TCs) in the *Drosophila* genetic screen. See text and [8] for additional details.

Supplementary Table 2

List of raw geometric features from the *Drosophila* screen

Geometric Feature (*Drosophila*)
Area
Solidity
Eccentricity
MajorAxisLength
MinorAxisLength
EquivDiameter
MeanIntensity
StdIntensity
90thPercentileIntensity
GFPBrightSpotMajorSegments
GFPBrightSpotTotalArea
GFPBrightSpotMajorSegmentAreaMean
GFPBrightSpotMajorSegmentAreaCV
GFPBrightSpotMajorSegmentMaxMinSeparation
GFPBrightSpotGFPCentroidRelOffset
GFPCentroidGFPCenterOfMassRelOffset
GFPBrightSpotGFPCenterOfMassRelOffset
GFPCenterOfMassGFPMomentOfInertia
GFPCentroidGFPMomentOfInertia
GFPBrightSpotGFPMomentOfInertia
GFPMultivariateKurtosis
GFPHalfMassRelDistanceFromBoundary
GFPHalfMassRelDistanceFromGFPCenterOfMass
GFPHalfMassRelDistanceFromGFPCentroid
GFPHalfMassRelDistanceFromGFPBrightSpotCentroid
RuffleArea
RufflePixSum
RuffleVolume
DrainageArea
DrainagePixSum
GFPEdgeNumber
GFPEdgeTotalPixels
GFPEdgePixelDensity
GFPEdgeMeanLength
GFPEdgeMeanRelativeLength
GFPIntensityLocationMutualInformation_5_15_15

GFPIntensityLocationMutualInformation_8_15_24
GFPIntensityLocationMutualInformation_5_20_15
GFPIntensityLocationMutualInformation_8_20_24
GFPGauss2DFitMeanResidual
GFPGauss2DFitCorrelation
GFPGauss2DFitRelativeSigmaRow
GFPGauss2DFitRelativeSigmaCol
GFPGauss2DFitRelativeOffsetMeanFromSegCentroid
GFPGauss2DFitRelativeOffsetMeanFromBrightSpotCentroid
LoSmoothEccentricity
LoSmoothMajorAxisLength
LoSmoothMinorAxisLength
LoSmoothEllipticity
LoSmoothGFPCentroidClosestFocusRelOffset
LoSmoothGFPCenterOfMassClosestFocusRelOffset
LoSmoothGFPBrightSpotClosestFocusRelOffset
LoSmoothBndNormIntegratedAbsAngle
LoSmoothBndUndulationCount
LoSmoothBndUndulationTotalRelativeArea
LoSmoothBndProcessesGE0.5
LoSmoothBndProcessesGE1
LoSmoothBndCurvatureSharpestProcess
LoSmoothAreaSharpestProcess
LoSmoothRelativeAreaSharpestProcess
LoSmoothBndCurvature2ndSharpestProcess
LoSmoothArea2ndSharpestProcess
LoSmoothRelativeArea2ndSharpestProcess
LoSmoothBndAngleSharpestProcessesGFPCentroid
LoSmoothBndAngleSharpestProcessesGFPCenterOfMass
LoSmoothBndAngleSharpestProcessesGFPBrightSpotCentroid
LoSmoothHeightTallestProcess
LoSmoothRelativeHeightTallestProcess
LoSmoothAreaTallestProcess
LoSmoothRelativeAreaTallestProcess
LoSmoothBaseTallestProcess
LoSmoothRelativeBaseTallestProcess
LoSmoothHeight2ndTallestProcess
LoSmoothRelativeHeight2ndTallestProcess
LoSmoothArea2ndTallestProcess
LoSmoothRelativeArea2ndTallestProcess
LoSmoothBase2ndTallestProcess

LoSmoothRelativeBase2ndTallestProcess
LoSmoothBndAngleTallestProcessesGFPCentroid
LoSmoothBndAngleTallestProcessesGFPCenterOfMass
LoSmoothBndAngleTallestProcessesGFPBrightSpotCentroid
LoSmoothBndLargestAreaForProcessGE0.5
LoSmoothBndLargestRelativeAreaForProcessGE0.5
LoSmoothBndSecondLargestAreaForProcessGE0.5
LoSmoothBndSecondLargestRelativeAreaForProcessGE0.5
LoSmoothBndAngleLargestProcessesGE0.5GFPCentroid
LoSmoothBndAngleLargestProcessesGE0.5GFPCenterOfMass
LoSmoothBndAngleLargestProcessesGE0.5GFPBrightSpotCentroid
LoSmoothBndLargestAreaForProcessGE1
LoSmoothBndLargestRelativeAreaForProcessGE1
LoSmoothBndSecondLargestAreaForProcessGE1
LoSmoothBndSecondLargestRelativeAreaForProcessGE1
LoSmoothBndAngleLargestProcessesGE1GFPCentroid
LoSmoothBndAngleLargestProcessesGE1GFPCenterOfMass
LoSmoothBndAngleLargestProcessesGE1GFPBrightSpotCentroid
HiSmoothEccentricity
HiSmoothMajorAxisLength
HiSmoothMinorAxisLength
HiSmoothEllipticity
HiSmoothGFPCentroidClosestFocusRelOffset
HiSmoothGFPCenterOfMassClosestFocusRelOffset
HiSmoothGFPBrightSpotClosestFocusRelOffset
HiSmoothBndNormIntegratedAbsAngle
HiSmoothBndUndulationCount
HiSmoothBndUndulationTotalRelativeArea
HiSmoothBndProcessesGE0.5
HiSmoothBndProcessesGE1
HiSmoothBndCurvatureSharpestProcess
HiSmoothAreaSharpestProcess
HiSmoothRelativeAreaSharpestProcess
HiSmoothBndCurvature2ndSharpestProcess
HiSmoothArea2ndSharpestProcess
HiSmoothRelativeArea2ndSharpestProcess
HiSmoothBndAngleSharpestProcessesGFPCentroid
HiSmoothBndAngleSharpestProcessesGFPCenterOfMass
HiSmoothBndAngleSharpestProcessesGFPBrightSpotCentroid
HiSmoothHeightTallestProcess
HiSmoothRelativeHeightTallestProcess

HiSmoothAreaTallestProcess
HiSmoothRelativeAreaTallestProcess
HiSmoothBaseTallestProcess
HiSmoothRelativeBaseTallestProcess
HiSmoothHeight2ndTallestProcess
HiSmoothRelativeHeight2ndTallestProcess
HiSmoothArea2ndTallestProcess
HiSmoothRelativeArea2ndTallestProcess
HiSmoothBase2ndTallestProcess
HiSmoothRelativeBase2ndTallestProcess
HiSmoothBndAngleTallestProcessesGFPCentroid
HiSmoothBndAngleTallestProcessesGFPCenterOfMass
HiSmoothBndAngleTallestProcessesGFPBrightSpotCentroid
HiSmoothBndLargestAreaForProcessGE0.5
HiSmoothBndLargestRelativeAreaForProcessGE0.5
HiSmoothBndSecondLargestAreaForProcessGE0.5
HiSmoothBndSecondLargestRelativeAreaForProcessGE0.5
HiSmoothBndAngleLargestProcessesGE0.5GFPCentroid
HiSmoothBndAngleLargestProcessesGE0.5GFPCenterOfMass
HiSmoothBndAngleLargestProcessesGE0.5GFPBrightSpotCentroid
HiSmoothBndLargestAreaForProcessGE1
HiSmoothBndLargestRelativeAreaForProcessGE1
HiSmoothBndSecondLargestAreaForProcessGE1
HiSmoothBndSecondLargestRelativeAreaForProcessGE1
HiSmoothBndAngleLargestProcessesGE1GFPCentroid
HiSmoothBndAngleLargestProcessesGE1GFPCenterOfMass
HiSmoothBndAngleLargestProcessesGE1GFPBrightSpotCentroid

Supplementary Table 2. Geometric features (145) measured in the *Drosophila* screen. See [8] for additional information on feature definitions.

Supplementary Table 3

List of raw geometric features from the yeast screen

Geometric Feature (yeast)	Used in Variability Analysis (yes/no)
x1	No
x2	No
y1	No
y2	No
Long axis end on neck side	No
Long axis end on hip side	No
Short axis end in mother cell on neck side	No
Short axis end in mother cell on hip side	No
Left neck position	No
Right neck position	No
Long axis end in bud on neck side	No
Logn axis end in bud on tip side	No
Short axis end in bud on left side	No
Short axis end in bud on right side	No
Brightest point on cell wall	No
Darkest point on cell wall	No
Farthest point from neck	No
Mother cell size	Yes
Area of daughter cell	Yes
Contour length of mother cell	Yes
Contour length of daughter cell	Yes
Fitness	Yes
Cell size	Yes
Contour length of cell	Yes
Long axis length of mother cell	Yes
Short axis length of mother cell	Yes
Neck position	Yes
Bud growth direction	Yes
Long axis length in bud	Yes
Short axis length in bud	Yes
Neck width	Yes
Length from bud tip to mother cell's long axis	Yes
Length from bud tip to mother cell's short axis	Yes
Distance from the center of the mother cell to its bud neck middle point	Yes
Distance from bud tip to mother cell's long axis along bud direction	Yes
Roundness of bud	Yes

Roundness of mother cell	Yes
Ratio of roundness of mother cell to that of bud	Yes
Ratio of the countour length	Yes
Ratio of the cell sizes	Yes
Length from bud neck to the farthest point on mother cell	Yes
Center of actin region in mother cell	No
Center of actin region in bud	No
Center of actin region	No
Center of actin patch in mother cell	No
Center of actin patch in bud	No
Center of actin patch	No
Center of gravity of actin region in mother cell	No
Center of gravity of actin region in bud	No
Center of gravity of actin region	No
Center of gravity of actin patch in mother cell	No
Center of gravity of actin patch in bud	No
Center of gravity of actin patch	No
Farthest point of actin region from neck in mother cell	No
Farthest point of actin region from neck in bud	No
Farthest point of actin region from neck	No
Actin region size in mother cell	Yes
Actin region size in bud	Yes
Ratio of actin region on neck	Yes
Actin region ratio	Yes
Actin region ratio in bud	Yes
Relative Distance of actin patch from neck in mother cell	Yes
Relative Distance of actin patch center from neck in bud	Yes
Total length of actin patch link	Yes
Maximum actin patch length	Yes
Number of actin patches	Yes
Ratio of actin patch region to actin region	Yes
Nucleus center in mother cell	No
Nucleus center in bud	No
Nucleus center in mother cell	No
Nucleus center in bud	No
D3-1	No
D3-2	No
D3-3	No
D4-1	No
D4-2	No
D4-3	No
D5-1	No
D5-2	No
D5-3	No
D6-1	No

D6-2	No
D7	No
D8	No
D9-1	No
D9-2	No
Mother nucleus border close to neck	No
Bud nucleus border close to neck	No
D12-1	No
D12-2	No
D13-1	No
D13-2	No
Area of nucleus region in mother cell	Yes
Area of nucleus region in bud	Yes
Area of nucleus region	Yes
Fitness to ellipse of the nucleus in the mother cell	Yes
Fitness to ellipse of the nucleus in the daughter cell	Yes
Fitness to ellipse of the nucleus	Yes
Number of nucleus	Yes
Distance from nuclear center to tip in unbudded cells	Yes
Distance from nuclear center to mother tip in budded cell	Yes
Distance from nuclear center to mother tip	Yes
Ratio of D102 to C103	Yes
Ratio of D103 to C103	Yes
Ratio of D104 to C103	Yes
Distance from neck to mother cell's nucleus	Yes
Distance from neck to bud's nucleus	Yes
Distance from neck to nucleus center	Yes
Distance from neck to nucleus center	Yes
Ratio_of_D108_to_C128_on_stage_C	Yes
Ratio_of_D109_to_C107	Yes
Ratio_of_D110_to_C128_on_stage_A1B	Yes
D115	Yes
Distance between two nucleus	Yes
Distance from mother cell's center to mother cell's nucleus	Yes
Distance from mother cell's center to mother cell's nucleus	Yes
Distance from bud center to bud's nucleus	Yes
Distance from bud center to nucleus center in A1B	Yes
Distance from bud nucleus to bud tip	Yes
Distance from nucleus to bud tip	Yes
Ratio_of_D121_to_C107	Yes
D124	Yes
Nucleus border point close to neck on mother cell's nucleus	Yes
Nucleus border point close to neck on bud's nucleus	Yes
Distance_between_nuclear_outline_point_C7_and_mother_hip_on_stage_A1B	Yes
Distance_between_nuclear_outline_point_C8_in_bud_and_bud_tip_on_stage_C	Yes

Relative_distance_of_nuclear_gravity_center_to_cell_center_on_stage_A	Yes
Relative_distance_of_nuclear_gravity_center_in_bud_to_bud_center_on_stage_C	Yes
Distance ratio of two nuclei from neck	Yes
Mobility of nucleus in mother cell	Yes
Mobility of nucleus in bud	Yes
Angle_between_C1D1-1_and_C1C1-2_on_stage_A	Yes
Angle_between_C2D1-2_and_C2C4-2_on_stage_C	Yes
Angle_between_D18-1D1-1_and_D18-1C1-2_on_stage_C	Yes
slope of mother nucleus	Yes
slope between two nuclei to neck position	Yes
slope of nuleus to neck position	Yes
slope between two nuclei to neck position	Yes
Angle_between_D18-2D1-2_and_D18-2C4-2_on_stage_C	Yes
D168	Yes
Angle_between_M1D1-1_and_M1C1_on_stage_A1B	Yes
Angle_between_M1D4_and_M1C1_on_stage_A1B	Yes
Angle_between_M1D4_and_M1C1_on_stage_A1B	Yes
nucleus maximum radius in mother cell	Yes
nucleusmaximum radius in bud	Yes
nucleus maximum radius	Yes
nucleus diameter in mother cell	Yes
nucleus diameter in bud	Yes
nucleus diameter	Yes
nucleus minimum radius in mother cell	Yes
nucleus minimum radius in bud	Yes
nucleus minimum radius	Yes
nucleus roundness in mother cell	Yes
nucleus roundness in bud	Yes
nucleus roundness	Yes
distance between nuclei through neck	Yes
distance between nuclei through neck	Yes
nuclei size ratio	Yes

Supplementary Table 3. Geometric features measured in the yeast screen. See [7] for feature definitions. The features used in our variability analysis are marked as such.

Supplementary Table 4

Principal components for the *Drosophila* screen

Geometric Feature (*Drosophila*)	PC1	PC2	PC3
Area	0.116	-0.186	0.018
Solidity	-0.142	-0.067	0.025
Eccentricity	0.086	0.127	0.118
MajorAxisLength	0.147	-0.028	0.073
MinorAxisLength	0.102	-0.196	-0.034
EquivDiameter	0.130	-0.175	0.031
MeanIntensity	-0.029	-0.029	-0.051
StdIntensity	0.016	-0.033	-0.057
90thPercentileIntensity	-0.008	-0.020	-0.059
GFPBrightSpotMajorSegments	0.026	-0.018	0.049
GFPBrightSpotTotalArea	0.113	-0.182	0.012
GFPBrightSpotMajorSegmentAreaMean	0.108	-0.174	0.000
GFPBrightSpotMajorSegmentAreaCV	0.027	-0.019	0.040
GFPBrightSpotMajorSegmentMaxMinSeparation	0.034	0.016	0.061
GFPBrightSpotGFPCentroidRelOffset	0.058	0.090	0.083
GFPCentroidGFPCenterOfMassRelOffset	0.061	0.090	0.071
GFPBrightSpotGFPCenterOfMassRelOffset	0.043	0.071	0.079
GFPCenterOfMassGFPMomentOfInertia	0.065	-0.078	0.035
GFPCentroidGFPMomentOfInertia	0.066	-0.076	0.036
GFPBrightSpotGFPMomentOfInertia	0.065	-0.076	0.036
GFPMultivariateKurtosis	0.094	0.046	-0.038
GFPHalfMassRelDistanceFromBoundary	-0.107	-0.084	-0.092
GFPHalfMassRelDistanceFromGFPCenterOfMass	0.018	0.056	0.084
GFPHalfMassRelDistanceFromGFPCentroid	0.048	0.087	0.093
GFPHalfMassRelDistanceFromGFPBrightSpotCentroid	-0.017	0.021	0.057
RuffleArea	0.131	-0.078	-0.040
RufflePixSum	0.115	-0.069	-0.030
RuffleVolume	0.049	0.030	-0.042
DrainageArea	0.080	-0.189	0.009
DrainagePixSum	0.080	-0.190	0.002
GFPEdgeNumber	0.115	-0.166	0.013
GFPEdgeTotalPixels	0.122	-0.163	0.006
GFPEdgePixelDensity	0.096	0.046	0.002
GFPEdgeMeanLength	-0.048	0.051	-0.059
GFPEdgeMeanRelativeLength	-0.082	0.071	-0.068
GFPIntensityLocationMutualInformation_5_15_15	0.026	-0.147	0.049
GFPIntensityLocationMutualInformation_8_15_24	0.051	-0.164	0.055
GFPIntensityLocationMutualInformation_5_20_15	0.054	-0.148	0.056
GFPIntensityLocationMutualInformation_8_20_24	0.072	-0.159	0.058

GFPGauss2DFitMeanResidual	0.011	0.026	0.012
GFPGauss2DFitCorrelation	0.002	-0.001	-0.004
GFPGauss2DFitRelativeSigmaRow	0.045	0.038	0.035
GFPGauss2DFitRelativeSigmaCol	0.075	0.060	-0.014
GFPGauss2DFitRelativeOffsetMeanFromSegCentroid	0.070	0.099	0.066
GFPGauss2DFitRelativeOffsetMeanFromBrightSpotCentroid	0.058	0.088	0.055
LoSmoothEccentricity	0.085	0.130	0.117
LoSmoothMajorAxisLength	0.147	-0.032	0.073
LoSmoothMinorAxisLength	0.102	-0.198	-0.029
LoSmoothEllipticity	0.002	0.002	0.004
LoSmoothGFPCentroidClosestFocusRelOffset	0.104	0.133	0.113
LoSmoothGFPCenterOfMassClosestFocusRelOffset	0.105	0.128	0.110
LoSmoothGFPBrightSpotClosestFocusRelOffset	0.103	0.121	0.100
LoSmoothBndNormIntegratedAbsAngle	0.139	-0.086	-0.069
LoSmoothBndUndulationCount	0.154	-0.087	-0.009
LoSmoothBndUndulationTotalRelativeArea	-0.133	0.073	-0.009
LoSmoothBndProcessesGE0.5	0.108	0.020	-0.141
LoSmoothBndProcessesGE1	0.093	0.036	-0.149
LoSmoothBndCurvatureSharpestProcess	0.004	0.008	-0.011
LoSmoothAreaSharpestProcess	-0.088	-0.040	-0.063
LoSmoothRelativeAreaSharpestProcess	-0.107	-0.016	-0.098
LoSmoothBndCurvature2ndSharpestProcess	0.050	0.024	-0.092
LoSmoothArea2ndSharpestProcess	-0.035	0.018	0.104
LoSmoothRelativeArea2ndSharpestProcess	-0.056	0.054	0.081
LoSmoothBndAngleSharpestProcessesGFPCentroid	0.042	0.090	0.111
LoSmoothBndAngleSharpestProcessesGFPCenterOfMass	0.043	0.089	0.109
LoSmoothBndAngleSharpestProcessesGFPBrightSpotCentroid	0.044	0.088	0.107
LoSmoothHeightTallestProcess	-0.088	-0.045	-0.112
LoSmoothRelativeHeightTallestProcess	-0.093	-0.040	-0.120
LoSmoothAreaTallestProcess	-0.087	-0.064	0.025
LoSmoothRelativeAreaTallestProcess	-0.135	0.002	-0.046
LoSmoothBaseTallestProcess	0.009	-0.047	0.111
LoSmoothRelativeBaseTallestProcess	-0.027	0.021	0.098
LoSmoothHeight2ndTallestProcess	0.057	0.032	0.104
LoSmoothRelativeHeight2ndTallestProcess	0.008	0.093	0.097
LoSmoothArea2ndTallestProcess	0.038	-0.006	0.109
LoSmoothRelativeArea2ndTallestProcess	-0.016	0.085	0.092
LoSmoothBase2ndTallestProcess	0.010	-0.047	0.111
LoSmoothRelativeBase2ndTallestProcess	-0.026	0.021	0.099
LoSmoothBndAngleTallestProcessesGFPCentroid	0.052	0.065	0.142
LoSmoothBndAngleTallestProcessesGFPCenterOfMass	0.052	0.064	0.142
LoSmoothBndAngleTallestProcessesGFPBrightSpotCentroid	0.051	0.062	0.139
LoSmoothBndLargestAreaForProcessGE0.5	0.052	0.039	-0.055
LoSmoothBndLargestRelativeAreaForProcessGE0.5	0.001	0.065	-0.036
LoSmoothBndSecondLargestAreaForProcessGE0.5	0.092	0.024	-0.138

LoSmoothBndSecondLargestRelativeAreaForProcessGE0.5	0.055	0.083	-0.111
LoSmoothBndAngleLargestProcessesGE0.5GFPCentroid	0.082	0.060	-0.124
LoSmoothBndAngleLargestProcessesGE0.5GFPCenterOfMass	0.083	0.059	-0.125
LoSmoothBndAngleLargestProcessesGE0.5GFPBrightSpotCentroid	0.083	0.059	-0.125
LoSmoothBndLargestAreaForProcessGE1	0.064	0.043	-0.089
LoSmoothBndLargestRelativeAreaForProcessGE1	0.020	0.071	-0.054
LoSmoothBndSecondLargestAreaForProcessGE1	0.076	0.028	-0.153
LoSmoothBndSecondLargestRelativeAreaForProcessGE1	0.050	0.069	-0.129
LoSmoothBndAngleLargestProcessesGE1GFPCentroid	0.069	0.052	-0.140
LoSmoothBndAngleLargestProcessesGE1GFPCenterOfMass	0.070	0.051	-0.141
LoSmoothBndAngleLargestProcessesGE1GFPBrightSpotCentroid	0.070	0.051	-0.141
HiSmoothEccentricity	0.063	0.146	0.091
HiSmoothMajorAxisLength	0.148	-0.038	0.073
HiSmoothMinorAxisLength	0.106	-0.193	-0.015
HiSmoothEllipticity	0.001	0.001	0.003
HiSmoothGFPCentroidClosestFocusRelOffset	0.115	0.115	0.110
HiSmoothGFPCenterOfMassClosestFocusRelOffset	0.113	0.108	0.105
HiSmoothGFPBrightSpotClosestFocusRelOffset	0.110	0.100	0.093
HiSmoothBndNormIntegratedAbsAngle	0.121	-0.061	-0.088
HiSmoothBndUndulationCount	0.153	-0.055	-0.057
HiSmoothBndUndulationTotalRelativeArea	-0.124	0.070	0.081
HiSmoothBndProcessesGE0.5	0.084	0.073	-0.147
HiSmoothBndProcessesGE1	0.070	0.075	-0.147
HiSmoothBndCurvatureSharpestProcess	0.002	0.001	-0.003
HiSmoothAreaSharpestProcess	-0.064	-0.077	0.085
HiSmoothRelativeAreaSharpestProcess	-0.141	0.001	0.025
HiSmoothBndCurvature2ndSharpestProcess	0.025	0.021	-0.070
HiSmoothArea2ndSharpestProcess	0.072	-0.033	0.076
HiSmoothRelativeArea2ndSharpestProcess	0.049	0.053	0.054
HiSmoothBndAngleSharpestProcessesGFPCentroid	0.126	0.056	0.021
HiSmoothBndAngleSharpestProcessesGFPCenterOfMass	0.126	0.057	0.020
HiSmoothBndAngleSharpestProcessesGFPBrightSpotCentroid	0.125	0.056	0.018
HiSmoothHeightTallestProcess	-0.106	-0.061	0.041
HiSmoothRelativeHeightTallestProcess	-0.132	-0.045	0.025
HiSmoothAreaTallestProcess	0.000	-0.141	0.095
HiSmoothRelativeAreaTallestProcess	-0.147	0.007	0.040
HiSmoothBaseTallestProcess	0.099	-0.010	-0.021
HiSmoothRelativeBaseTallestProcess	0.074	0.058	-0.029
HiSmoothHeight2ndTallestProcess	0.097	-0.015	0.074
HiSmoothRelativeHeight2ndTallestProcess	0.074	0.039	0.069
HiSmoothArea2ndTallestProcess	0.105	-0.064	0.055
HiSmoothRelativeArea2ndTallestProcess	0.085	0.062	0.041
HiSmoothBase2ndTallestProcess	0.100	-0.010	-0.021
HiSmoothRelativeBase2ndTallestProcess	0.075	0.058	-0.028
HiSmoothBndAngleTallestProcessesGFPCentroid	0.131	0.038	0.027

HiSmoothBndAngleTallestProcessesGFPCenterOfMass	0.131	0.039	0.025
HiSmoothBndAngleTallestProcessesGFPBrightSpotCentroid	0.131	0.038	0.023
HiSmoothBndLargestAreaForProcessGE0.5	0.019	0.064	-0.048
HiSmoothBndLargestRelativeAreaForProcessGE0.5	-0.020	0.080	-0.048
HiSmoothBndSecondLargestAreaForProcessGE0.5	0.062	0.041	-0.118
HiSmoothBndSecondLargestRelativeAreaForProcessGE0.5	0.035	0.068	-0.096
HiSmoothBndAngleLargestProcessesGE0.5GFPCentroid	0.063	0.073	-0.139
HiSmoothBndAngleLargestProcessesGE0.5GFPCenterOfMass	0.063	0.073	-0.139
HiSmoothBndAngleLargestProcessesGE0.5GFPBrightSpotCentroid	0.063	0.073	-0.139
HiSmoothBndLargestAreaForProcessGE1	0.018	0.061	-0.064
HiSmoothBndLargestRelativeAreaForProcessGE1	-0.015	0.070	-0.056
HiSmoothBndSecondLargestAreaForProcessGE1	0.050	0.036	-0.121
HiSmoothBndSecondLargestRelativeAreaForProcessGE1	0.026	0.054	-0.097
HiSmoothBndAngleLargestProcessesGE1GFPCentroid	0.050	0.061	-0.137
HiSmoothBndAngleLargestProcessesGE1GFPCenterOfMass	0.050	0.061	-0.137
HiSmoothBndAngleLargestProcessesGE1GFPBrightSpotCentroid	0.049	0.061	-0.137

Supplementary Table 4. The coordinates for the first three principal components for the *Drosophila* screen are given here. The first column contains the list of 145 raw features. The next three columns contain the coordinates for the first three PCs.

Supplementary Table 5

Principal components for the yeast screen

Geometric Feature (yeast)	PC1	PC2	PC3
Mother cell size	0.265	-0.150	0.161
Area of daughter cell	0.146	0.206	-0.039
Contour length of mother cell	0.183	-0.146	0.158
Contour length of daughter cell	0.141	0.210	-0.048
Fitness	0.000	0.000	0.000
Cell size	0.306	0.032	0.075
Contour length of cell	0.284	0.061	0.047
Long axis length of mother cell	0.255	-0.142	0.167
Short axis length of mother cell	0.242	-0.142	0.138
Neck position	0.026	0.001	0.012
Bud growth direction	0.021	0.001	0.018
Long axis length in bud	0.141	0.205	-0.047
Short axis length in bud	0.122	0.173	-0.031
Neck width	0.063	0.017	0.019
Length from bud tip to mother cell's long axis	0.101	0.108	-0.006
Length from bud tip to mother cell's short axis	0.057	0.082	-0.023
Distance from the center of the mother cell to its bud neck middle point	0.121	-0.054	0.089
Distance from bud tip to mother cell's long axis along bud direction	0.050	0.057	-0.009
Roundness of bud	0.071	0.088	-0.021
Roundness of mother cell	0.013	0.004	0.038
Ratio of roundness of mother cell to that of bud	0.060	0.087	-0.037
Ratio of the countour length	0.104	0.232	-0.076
Ratio of the cell sizes	0.103	0.235	-0.073
Length from bud neck to the farthest point on mother cell	0.129	-0.066	0.092
Actin region size in mother cell	-0.007	0.054	0.007
Actin region size in bud	0.101	0.078	-0.032
Ratio of actin region on neck	-0.025	-0.027	0.030
Actin region ratio	-0.018	0.085	-0.047
Actin region ratio in bud	0.037	0.004	-0.034
Relative Distance of actin patch from neck in mother cell	0.003	0.004	-0.005
Relative Distance of actin patch center from neck in bud	0.033	0.014	-0.026
Total length of actin patch link	0.188	0.084	-0.057
Maximum actin patch length	0.185	0.082	-0.050
Number of actin patches	0.107	0.003	-0.058
Ratio of actin patch region to actin region	0.010	-0.029	-0.053
Area of nucleus region in mother cell	0.179	-0.228	-0.274
Area of nucleus region in bud	0.030	0.019	0.013
Area of nucleus region	0.259	-0.006	-0.242
Fitness to ellipse of the nucleus in the mother cell	0.000	0.000	0.000
Fitness to ellipse of the nucleus in the daughter cell	0.000	0.000	0.000
Fitness to ellipse of the nucleus	0.000	0.000	0.000
Number of nucleus	0.164	0.271	-0.017

Distance from nuclear center to tip in unbudded cells	0.078	-0.063	0.139
Distance from nuclear center to mother tip in budded cell	0.024	-0.196	0.089
Distance from nuclear center to mother tip	0.163	-0.140	0.225
Ratio of D102 to C103	0.019	-0.007	0.140
Ratio of D103 to C103	0.000	0.000	0.000
Ratio of D104 to C103	0.000	0.000	0.000
Distance from neck to mother cell's nucleus	0.084	0.196	0.000
Distance from neck to bud's nucleus	0.015	0.058	0.065
Distance from neck to nucleus center	0.005	-0.022	-0.020
Distance from neck to nucleus center	0.000	0.000	0.000
Ratio_of_D108_to_C128_on_stage_C	0.051	0.221	-0.024
Ratio_of_D109_to_C107	-0.002	0.057	0.061
Ratio_of_D110_to_C128_on_stage_A1B	-0.028	0.001	-0.046
D115	0.000	0.000	0.000
Distance between two nucleus	0.018	0.068	0.072
Distance from mother cell's center to mother cell's nucleus	0.147	0.015	0.349
Distance from mother cell's center to mother cell's nucleus	0.115	-0.074	0.268
Distance from bud center to bud's nucleus	0.018	-0.027	-0.013
Distance from bud center to nucleus center in A1B	0.003	-0.001	-0.007
Distance from bud nucleus to bud tip	0.022	-0.048	-0.043
Distance from nucleus to bud tip	0.006	0.000	-0.002
Ratio_of_D121_to_C107	0.001	-0.058	-0.059
D124	-0.002	-0.001	-0.007
Nucleus border point close to neck on mother cell's nucleus	0.065	0.190	0.026
Nucleus border point close to neck on bud's nucleus	0.016	0.026	0.039
Distance_between_nuclear_outline_point_C7_and_mother_hip_on_stage_A1B	-0.009	-0.193	0.086
Distance_between_nuclear_outline_point_C8_in_bud_and_bud_tip_on_stage_C	0.010	-0.021	-0.025
Relative_distance_of_nuclear_gravity_center_to_cell_center_on_stage_A	0.079	0.072	0.294
Relative_distance_of_nuclear_gravity_center_in_bud_to_bud_center_on_stage_C	0.004	0.006	0.010
Distance ratio of two nuclei from neck	0.005	-0.013	-0.008
Mobility of nucleus in mother cell	0.047	0.181	0.019
Mobility of nucleus in bud	0.004	0.024	0.032
Angle_between_C1D1-1_and_C1C1-2_on_stage_A	-0.013	0.058	-0.063
Angle_between_C2D1-2_and_C2C4-2_on_stage_C	-0.004	0.015	0.012
Angle_between_D18-1D1-1_and_D18-1C1-2_on_stage_C	0.007	-0.008	0.010
slope of mother nucleus	0.000	-0.004	0.013
slope between two nuclei to neck position	0.001	-0.027	-0.001
slope of nuleus to neck position	-0.004	-0.003	0.027
slope between two nuclei to neck position	0.000	-0.001	0.007
Angle_between_D18-2D1-2_and_D18-2C4-2_on_stage_C	0.001	0.007	0.008
D168	0.000	0.000	0.003
Angle_between_M1D1-1_and_M1C1_on_stage_A1B	-0.007	-0.071	0.097
Angle_between_M1D4_and_M1C1_on_stage_A1B	0.014	-0.007	0.020
Angle_between_M1D4_and_M1C1_on_stage_A1B	0.000	-0.001	0.006
nucleus maximum radius in mother cell	0.163	-0.166	-0.264
nucleusmaximum radius in bud	0.024	-0.008	-0.015

nucleus maximum radius	0.144	-0.093	-0.223
nucleus diameter in mother cell	0.166	-0.174	-0.264
nucleus diameter in bud	0.026	-0.008	-0.016
nucleus diameter	0.146	-0.094	-0.223
nucleus minimum radius in mother cell	0.113	-0.217	-0.153
nucleus minimum radius in bud	0.021	0.005	0.001
nucleus minimum radius	0.106	-0.148	-0.108
nucleus roundness in mother cell	0.000	0.000	0.000
nucleus roundness in bud	0.000	0.000	0.000
nucleus roundness	0.000	0.000	0.000
distance between nuclei through neck	0.049	0.020	0.038
distance between nuclei through neck	0.007	-0.001	0.004
nuclei size ratio	0.002	0.015	0.025

Supplementary Table 5. The coordinates for the first three principal components for the yeast screen are given here. The first column contains the list of 101 raw features that were used in the variability analysis. The next three columns contain the coordinates for the first three PCs.

Supplementary Table 6

Analysis of robustness to method of dimensionality reduction for *Drosophila* TCs with low variability p-scores

Method	1 PC	2 PCs	3 PCs	5 PCs	10 PCs	NNs
1 PC	-	<1E-14	7.73E-14	2.88E-14	<1E-14	3.06E-11
2 PCs	<1E-14	-	8.80E-14	5.32E-14	<1E-14	3.38E-13
3 PCs	7.73E-14	8.80E-14	-	7.16E-14	<1E-14	1.92E-12
5 PCs	2.88E-14	5.32E-14	7.16E-14	-	<1E-14	<1E-14
10 PCs	<1E-14	<1E-14	<1E-14	<1E-14	-	5.76E-13
NNs	3.06E-11	3.38E-13	1.92E-12	<1E-14	5.76E-13	--

Supplementary Table 6. Robustness of variability p-scores to method used for dimensionality reduction. We computed variability p-scores for all *Drosophila* TCs using six different methods for dimensionality reduction. The first five methods were to use 1, 2, 3, 5, or 10 principal components, while the sixth method was to use the best 7 neural network classifiers constructed in [8]. For each method, we calculated variability p-scores using bootstrapping with 10^3 iterations. We then considered the set of TCs which had variability p-scores $< 10^{-3}$ (i.e., TCs for which none of the bootstrapped samples had a lower variability v-score) for each method. We used the hypergeometric cdf to calculate the probability of the observed overlap between these sets for each pair of the six alternate dimensionality reduction methods. The probabilities of overlap are all $< 10^{-10}$, indicating a high degree of invariance in the ordering of TCs at the lower end of variability p-scores, despite alterations in the method of dimensionality reduction.

Supplementary Table 7

Analysis of robustness to method of dimensionality reduction for *Drosophila* TCs with high variability p-scores

Method	1 PC	2 PCs	3 PCs	5 PCs	10 PCs	NNs
1 PC	-	3.15E-12	6.74E-10	6.74E-10	3.15E-12	1.75E-04
2 PCs	3.15E-12	-	3.15E-12	6.74E-10	6.74E-10	1.75E-04
3 PCs	6.74E-10	3.15E-12	-	3.15E-12	3.15E-12	1.75E-04
5 PCs	6.74E-10	6.74E-10	3.15E-12	-	2.50E-13	1.75E-04
10 PCs	3.15E-12	6.74E-10	3.15E-12	2.50E-13	-	1.75E-04
NNs	1.75E-04	1.75E-04	1.75E-04	1.75E-04	1.75E-04	-

Supplementary Table 7. Robustness of variability p-scores to method used for dimensionality reduction. We computed variability p-scores for all *Drosophila* TCs using six different methods for dimensionality reduction. The first five methods were to use 1, 2, 3, 5, or 10 principal components, while the sixth method was to use the best 7 neural network classifiers constructed in [8]. For each method, we calculated variability p-scores using bootstrapping with 10^3 iterations. We then considered the set of 10 top-scoring TCs for each method. We used the hypergeometric cdf to calculate the probability of the observed overlap between these sets for each pair of the six alternate dimensionality reduction methods. Except for the case of the neural network method vs. a PC-based method, the probabilities of pairwise overlap are all $< 10^{-9}$, indicating a high degree of invariance in the ordering of TCs at the higher end of variability p-scores, despite alterations in the method of dimensionality reduction. See the text for an explanation of the relatively large deviation in the top-scoring TCs for the neural network method versus the various PC-based methods of dimensionality reduction. In short, this is due to the fact that a neural network dimensionality reduction is likely to lose more information from the raw feature data than a PC-based method.

Supplementary Table 8

Standard errors for variability p-scores in *Drosophila* from jackknifing

Treatment Condition	Variability p-Score	Standard Error
Goalpha65A_overex'	0	0
Graf:Graf_RNAi__P1P7'	0	0
gartenzwerg'	0	0.0008
Ankyrin'	0	0.0001
C3G'	0	0
CG9426'	0	0
CG9699'	0	0
CG4448'	0	0
CG7578'	0	0
CG12102'	0	0
Rho1'	0	0
twinstar'	0	0
Gef26'	0	0
Arf51F'	0	0
RacF28L'	0	0.0002
RacV12'	0	0.0003
RhoF30L'	0	0
apc2'	0	0
delRhoGEF3_const_overexp'	0	0
oncoSif+MTL_RNAi'	0	0
oncoSif+Rho1_RNAi'	0	0
oncosif_control'	0	0.0001
Merlin'	0.0001	0
CG13692'	0	0.0009
rho-like'	0	0.0004
cenB1A'	0	0.0004
RacGAP50C'	0.0002	0
Sif1_full_overexp'	0	0.0011
Trio:Trio_RNAi__P1B4'	0.0003	0
apc'	0.0003	0
armadillo'	0.0004	0
CG15611:P1M10__P1P19'	0	0.0014
shibire '	0	0.0005
G-gamma30A:P1N15__P1M15'	0.0002	0.0005
GefG64C_overex'	0.0008	0.0001
GEF64C:GEF64C_v361__GEF64C_08319__GEF64C_08318'	0.0012	0
cappuccino'	0	0.0015
CG18858 '	0	0.0018

CG4853 '	0.0018	0
G protein s60A '	0.0004	0.0016
CG5337'	0.0019	0.0006
Rab35'	0.0006	0.0024
staufen'	0.003	0.0001
Marf'	0.0029	0.0009
Rab9'	0.0043	0
CG12736'	0	0.0072
Actin-related protein 66B'	0.0008	0.0037
no receptor potential A'	0	0.0045
paxillin:P1M19__P1C11'	0	0.0062
dia'	0.0047	0.0005
CSN1a'	0.0043	0.0012
empty:P1F9__P1I23'	0.0047	0.0008
RhoBTB'	0.0035	0.0021
Msp-300'	0.0048	0.0008
kelch'	0.0028	0.0028
l(1)dd4'	0.0052	0.0004
mini spindles '	0.0007	0.005
Rab-protein 3 '	0.002	0.0039
hAuroraB_CA'	0.0061	0.0004
sec23'	0.0065	0.0009
CG9243'	0.0027	0.0049
CG32627'	0.0049	0.0038
CG8707 '	0.0009	0.0081
Rab26'	0.006	0.0031
capping protein beta '	0.007	0.0022
oncoSif+Rac1_RNAi'	0.0089	0.0003
abnormal spindle '	0.0092	0.0001
empty'	0	0.0279
oncosif_EnadsRNA'	0.0066	0.0026
mbc:mbc_16995__mbc_36492'	0.0084	0.0011
cenG1A:P1O17__P1P17'	0.001	0.0087
Arc-p20'	0.008	0.002
Microtubule-associated protein 205'	0.0021	0.0086
cnn'	0.0103	0.0013
RhoGEF3:P1O16__P1E2'	0.0132	0
ADP ribosylation factor 79F '	0.0019	0.0116
MICAL-like '	0.0122	0.0021
TumL'	0.0145	0.0003
oncosif_RhoGAP16FdsRNA'	0	0.0223
Mapmodulin'	0.0014	0.0137
Arc-p34'	0	0.0155
RhoGAP19B_RNAi'	0.0119	0.0035
falten '	0	0.0221

oncosif_RhoGAP54DdsRNA'	0	0.0381
Patsas '	0.0128	0.0039
RhoGAP68F '	0	0.0327
CG3009'	0	0.0283
sif'	0.0139	0.0046
Grip75'	0.0171	0.0019
RanGAP'	0.0149	0.0045
Rab-protein 6 '	0.0138	0.0061
CG8243:P1I14__P1O6'	0.0152	0.0049
GEF64C_25939'	0.0172	0.0053
CG32030 '	0.0184	0.0061
SCAR'	0.0169	0.0078
CdGAPr:CdGAPr_RNAi__P1I2'	0.0242	0.0013
CG11063'	0.0153	0.0107
dLis1_overex'	0.0262	0.0002
Brahma associated protein 55kD'	0	0.0384
CG5522 '	0.0222	0.0065
ran-like '	0.0061	0.0243
Septin-2'	0.0241	0.0079
CG7787'	0.0257	0.0081
Actn'	0.0131	0.0208
CG8397 '	0.0116	0.0237
alpha-Catenin'	0.0219	0.014
RhoGAP71E '	0.0121	0.0257
chrowded'	0.03	0.0082
jitterbug:P1F13__P1I4'	0.0293	0.0092
CG33232 '	0	0.0948
CG11968'	0.0338	0.0061
Rheb '	0.0238	0.0173
CG5160'	0.0376	0.004
CG1193 '	0.0304	0.0118
par1a1'	0.0416	0.001
GXIVsPLA2'	0.0418	0.0021
jitterbug'	0	0.0524
CG30440 '	0.0293	0.0162
CG7846'	0.038	0.0111
G protein 1 '	0.0039	0.048
lightoid'	0.0407	0.0117
rab3-GAP '	0.0491	0.0039
Bj1 protein '	0.0327	0.0206
Neurofibromin 1'	0.0521	0.0012
Sar1:P1M14__P1E20'	0.0479	0.0072
EG:BACH7M4.1_RNAi'	0.0201	0.0354
Fimbrin'	0.0524	0.0035
CG1583'	0.0562	0.0027

CG15609 '	0.0054	0.0547
CG4267'	0.0099	0.052
CG10724 '	0.0408	0.0219
gfp_06Oct17'	0.0246	0.0381
MTL_16751'	0.0378	0.0254
cib'	0.0627	0.0007
CG10540 '	0.0589	0.0055
CG9248'	0.056	0.0094
RhoGAPp190:RhoGAPp190_RNAi__P1O9'	0.0653	0.002
Menin 1'	0.0203	0.0474
oncosif_Rab5dsRNA'	0.0579	0.0117
CG14034'	0.0581	0.0172
capt'	0.058	0.0216
RapGAP1'	0.069	0.0118
RabX4 '	0.0668	0.0143
Rap21'	0.0636	0.0177
oncosif_RhoGEF3dsRNA'	0.0297	0.0556
CG8801 '	0.0667	0.0196
CLIP-190'	0.0573	0.0296
yurt '	0.0846	0.0045
CG15097 '	0.0201	0.0693
Centrosomal protein 190kD '	0.0905	0.0083
Crag'	0.0143	0.0852
RhoGAP5A_RNAi'	0.083	0.0167
vb'	0.0812	0.0198
CG32138'	0.0983	0.0047
CG7940 '	0.0997	0.0052
pod1'	0.0959	0.0282
Muscle-specific protein 300'	0.1159	0.0294
rtGEF_RNAi'	0.0906	0.0568
Elongation factor 148D '	0.1153	0.0335
Dia_RNAi'	0.1028	0.0502
Rac1_RNAi'	0.1391	0.0162
RhoGAP15B:RhoGAP15B_RNAi__P1M9'	0.1546	0.0183
enabled'	0.1236	0.0507
Cdep '	0.1323	0.0434
Sop2'	0.1656	0.0106
G protein -subunit 76C '	0.1588	0.0192
gamma-tubulin ring protein 84 '	0.1483	0.0321
CG6017 '	0.1515	0.0318
CG7420 '	0.1579	0.0258
par1_overex'	0.1724	0.0168
CG30440:CG30440_RNAi__P1J8'	0.1662	0.0352
CG6838'	0.1901	0.0363
RhoGEF3_RNAi'	0.1843	0.0437

Rab-protein 7 '	0.154	0.0816
oncosif_Arcp34dsRNA'	0.1961	0.0422
RhoGAP54D'	0.2089	0.0316
CG7197 '	0.2199	0.0292
Cdc42Y32A'	0.2456	0.0041
p16-ARC'	0.1991	0.0596
RacGAP50C_RNAi'	0.2294	0.0318
CG30456:P1L10__P1O18'	0.224	0.0429
EfSec'	0.2221	0.0471
CG7365 '	0.1355	0.1362
centaurin gamma 1A '	0.2054	0.0715
canoe '	0.2603	0.0478
RhoGAP1A_RNAi'	0.2784	0.03
CG16728'	0.2445	0.0668
memo'	0.2865	0.0404
gfp1'	0.3231	0.0265
RhoGEF3'	0.3291	0.0213
Rab-protein 8'	0.337	0.053
oncoSif'	0.3584	0.0529
Pld'	0.3613	0.0622
RabX2'	0.3405	0.0844
CG33275'	0.3542	0.0742
Rab30'	0.3732	0.0678
RacGAP50C:RacGAP50C_33345__RacGAP50C_07575'	0.3501	0.1058
Vav_RNAi'	0.4303	0.0359
gTub37C'	0.378	0.0957
miranda '	0.4229	0.0664
Microtubule-associated protein 60 '	0.3985	0.1044
RhoGAP16F:RhoGAP16F_RNAi__P1I11'	0.4539	0.0497
RhoGAP100F '	0.379	0.1291
CG14782'	0.4606	0.0546
MTL_36493'	0.4625	0.067
pebble '	0.4165	0.1184
CG15611_RNAi'	0.4881	0.0527
sanpodo '	0.541	0
RhoGAP18B_RNAi'	0.4973	0.0535
homolog of RecQ '	0.5046	0.0532
Rac1_Rac2_MTL_RNAi'	0.5137	0.0549
CG5022'	0.5414	0.036
peanut '	0.5214	0.0594
Sif1_RNAi'	0.5839	0.0051
Spectrin'	0.5689	0.0386
Dystrobrevin-like '	0.5382	0.0695
Moesin'	0.5737	0.0563
RhoGAP19D'	0.5936	0.0367

CG11490:P1B1__P1B3'	0.5981	0.0324
RhoGAP92B_RNAi'	0.6076	0.0295
nrg'	0.6149	0.0228
locomotion defects'	0.5789	0.0602
Rab-protein 2 '	0.5942	0.0509
G protein -subunit 13F '	0.5601	0.0889
CG30115:CG30115_RNAi__P1N21'	0.6218	0.0331
CG7323_RNAi'	0.5883	0.0724
oncosif_CG3799dsRNA'	0.6327	0.0306
CG3799_overexp'	0.6274	0.0495
formin 3'	0.632	0.0558
Grp1'	0.6406	0.0585
CG10971'	0.6507	0.0548
CG30372:P1D17__P1N6'	0.6549	0.0526
CG3799:CG3799_RNAi__P1G10'	0.6891	0.0559
CG14507'	0.6867	0.0649
concertina '	0.6814	0.0713
strad'	0.7389	0.0226
Gelsolin'	0.7381	0.0253
Rab-RP4'	0.7407	0.0269
alpha-cat related'	0.7029	0.0699
Sos:Sos_RNAi__P1N17'	0.7598	0.0382
G protein i subunit 65A'	0.7401	0.092
RhoV14'	0.8007	0.0323
Mp20'	0.7975	0.0408
CG12241'	0.7748	0.0664
CG31683 '	0.7615	0.0811
RhoGEF4:P1F6__RhoGEF4_11011'	0.8002	0.0604
G protein o47A '	0.8153	0.0485
CG12102:P1N16__P1I12'	0.8467	0.0202
RhoGEF2:RhoGEF2_07531__RhoGEF2_29373'	0.8374	0.0312
RhoGAP93B_RNAi'	0.8437	0.0365
moodybeta'	0.8527	0.0323
visceral mesodermal armadillo-repeats '	0.843	0.0427
empty'	0.8676	0.0189
Phospholipase A2 activator protein '	0.8491	0.047
Rab5'	0.863	0.0342
lava lamp '	0.8701	0.0282
CG10188:CG10188_RNAi__P1D11'	0.8727	0.0266
Cdep_RNAi'	0.8846	0.018
RhoGAP102A_RNAi'	0.8526	0.0564
CG9135/RCC-1'	0.8375	0.0731
CG8479'	0.894	0.0172
CG5745'	0.8832	0.0284
-Tubulin at 23C'	0.8562	0.0587

CG30456_RNAi'	0.8988	0.0246
control1'	0.9237	0.0074
Septin-5'	0.9205	0.0186
CG7324 '	0.947	0.0054
CG8557_RNAi'	0.951	0.0039
Rho1:Rho1_RNAi__P1F21__P1J16'	0.9608	0.0066
CG30158:P1K6__P1M6'	0.9756	0.0004
Cdc42_RNAi'	0.9873	0.0021
G protein 49B'	0.9536	0.0381
pbl:pbl_33336__pbl_11381__pbl_26301__pbl_RNAi__pbl_33335'	1	0

Supplementary Table 8. Jackknife statistics for all TCs in the *Drosophila* screen were computed, which allowed us to calculate standard errors for each TC's variability v-score and variability p-score. TCs were sorted according to variability p-score, and standard errors are as shown.

Supplementary Table 9

Variability p-scores for *Drosophila* phenocluster for lamellipodia formation

Treatment Condition	Variability p-Score
Trio:Trio_RNAi__P1B4	0.0157
gartenzwerg	0.0527
twinstar	0.0686
RhoF30L	0.0689
Rab26	0.1113
Grip75	0.1193
Graf:Graf_RNAi__P1P7	0.1323
rab3-GAP	0.1412
GEF64C:GEF64C_v361__GEF64C_08319__GEF64C_08318	0.2510
alpha-Catenin	0.2610
Septin-2	0.2617
kelch	0.2734
CSN1a	0.3029
GEF64C_25939	0.3476
CG15097	0.3481
CG32030	0.3612
Marf	0.3667
CG15611:P1M10__P1P19	0.4018
Arc-p20	0.4044
RhoGAP68F	0.4138
Rab35	0.4167
Mapmodulin	0.5106
CG3009	0.5619
RanGAP	0.5818
cnn	0.6294
capt	0.9306
CG15611_RNAi	0.9788

Supplementary Table 9. Variability p-scores for TCs in the lamellipodia formation phenocluster, as identified in [8]. The variability v-scores (v_i) were calculated as defined in the text, and the variability p-scores were calculated using a slightly modified procedure, where the bootstrapping sampling is taken from the set of cells comprising the TCs in the phenocluster, rather than all TCs in the genetic screen (see **Materials and Methods** for details). After Bonferroni correction, no TCs displayed significantly decreased or increased population

variability.

Supplementary Table 10

Variability p-scores for *Drosophila* phenocluster for protrusion/adhesion formation

Treatment Condition	Variability p-Score
Goalpha65A_overex	0.0000
Gef26	0.0000
delRhoGEF3_const_overexp	0.0000
Arf51F	0.0004
CG4448	0.0010
CG9699	0.0015
C3G	0.0151
CG9426	0.0157
Merlin	0.0179
rho-like	0.0236
Ankyrin	0.0665
CG4853	0.1030
Armadillo	0.1754
CG7578	0.1941
staufen	0.2088
Rab9	0.2096
l(1)dd4	0.2244
Centrosomal protein 190kD	0.2275
CG10540	0.2337
Sop2	0.2712
CG33232	0.2830
CG6838	0.3080
CG4267	0.4270
CG5337	0.4848
CG8801	0.5300
CG9248	0.5465
SCAR	0.5884
lightoid	0.5896
cib	0.5996
CG5160	0.6075
RapGAP1	0.6484
Rab30	0.8133
CG1583	0.8851
CG7846	0.9261
RhoGEF3	0.9598
RhoGAPp190	0.9780

Supplementary Table 10. Variability p-scores for TCs in the protrusion/adhesion formation phenocluster, as identified in [8]. The variability v-scores (v_i) were calculated as defined in the text, and the variability p-scores were calculated using a slightly modified procedure, where the bootstrapping sampling is taken from the set of cells comprising the TCs in the phenocluster, rather than all TCs in the genetic screen (see **Materials and Methods** for details). After Bonferroni correction, Goalpha65A_overex, Gef26, delRhoGEF3_const_overexp, Arf51F, and CG4448 have significantly reduced population variability ($p < .05/36$). No TCs had significantly increased population variability after Bonferroni-correction, although RhoGEF3 and RhoGAPp190 had marginally increased population variability (see main text for further discussion).

Supplementary Table 11

Variability p-scores for *Drosophila* phenocluster for adhesion disassembly/cortical tension

Treatment Condition	Variability p-Score
RacGAP50C	0.0000
oncoSif+MTL_RNAi	0.0000
CG12102	0.0001
RacV12	0.0001
oncosif_control	0.0002
RacF28L	0.0003
cenB1A	0.0005
CG13692	0.0012
CG18858	0.0023
Sif1_full_overexp	0.0026
MICAL-like	0.0037
CG9243	0.0050
CG5522	0.0077
Patsas	0.0093
CG12736	0.0116
paxillin:P1M19__P1C11	0.0121
Actin-related protein 66B	0.0136
mini spindles	0.0141
Sar1:P1M14__P1E20	0.0166
ADP ribosylation factor 79F	0.0213
CG11063	0.0228
Arc-p34	0.0278
oncosif_Rab5dsRNA	0.0319
CG30440	0.0380
oncosif_RhoGAP16FdsRNA	0.0528
CG10724	0.0547
Actn	0.0588
falten	0.0682
Rap21	0.0699
RhoGAP15B:RhoGAP15B_RNAi__P1M9	0.0700
Brahma associated protein 55kD	0.0827
oncosif_RhoGAP54DdsRNA	0.0934
gamma-tubulin ring protein 84	0.0960
RhoGAP92B_RNAi	0.1152
Menin 1	0.1373
alpha-cat related	0.1955
Crag	0.1998
CG15609	0.2193
pod1	0.2367

centaurin gamma 1A	0.2518
RhoGAP54D	0.2608
CG30440:CG30440_RNAi__P1J8	0.3252
CG16728	0.3319
G protein i subunit 65A	0.3348
CG33275	0.3459
RhoGAP19D	0.3755
CG30456:P1L10__P1O18	0.3897
EfSec	0.3955
CG7365	0.4013
Moesin	0.4352
sanpodo	0.4808
CG14782	0.5087
gTub37C	0.5233
Dystrobrevin-like	0.5293
RhoGAP100F	0.5972
Grp1	0.6090
oncosif_CG3799dsRNA	0.6233
locomotion defects	0.6240
G protein -subunit 13F	0.6364
CG31683	0.6543
CG10971	0.6604
Rab-protein 2	0.6663
Rab-RP4	0.6678
CG12102:P1N16__P1I12	0.6897
concertina	0.7122
Mp20	0.7433
Phospholipase A2 activator protein	0.7569
-Tubulin at 23C	0.7601
CG9135/RCC-1	0.7618
Gelsolin	0.7782
CG30456_RNAi	0.7850
G protein 49B	0.7967
Rab5	0.8608
lava lamp	0.8661
CG8479	0.8662
Septin-5	0.9023
CG30158:P1K6__P1M6	0.9759
pbl:pbl_33336__pbl_11381__pbl_26301__pbl_RNAi__pbl_33335	1.0000

Supplementary Table 11. Variability p-scores for TCs in the adhesion disassembly/cortical tension phenocluster, as identified in [8]. The variability v-scores (v_i) were calculated as defined in the text, and the variability p-scores were calculated using a slightly modified procedure,

where the bootstrapping sampling is taken from the set of cells comprising the TCs in the phenocluster, rather than all TCs in the genetic screen (see **Materials and Methods** for details). After Bonferroni correction, RacGAP50C, oncoSif+MTL_RNAi, RacV12, oncosif_control, RacF28L, and cenB1A have significantly reduced population variability ($p < .05/78$). Note that CG12102 had significantly decreased variability as one TC, but not in a duplicate, so it is not included in the final list. A single TC, pbl, had significantly increased population variability ($p > 1 - .05/78$).

Supplementary Table 12

Variability p-scores and percentile ranks for yeast TCs involved in septin ring recruitment and assembly

Treatment Condition	Variability p-Score	Percentile Rank	Gene Function
CDC3	N/A	N/A	Septin
CDC10	0.99989	99.96	Septin
CDC11	N/A	N/A	Septin
CDC12	N/A	N/A	Septin
SHS1	0.99963	97.83	Septin
CDC42	N/A	N/A	Activity required for septin recruitment and assembly
CDC24	N/A	N/A	CDC42 GEF
BEM3	0.0267	58.70	CDC42 GAP
RGA1	0.0628	64.19	CDC42 GAP
RGA2	4.06E-4	36.39	CDC42 GAP
CLA4	1.000	100	Regulates organization of septin ring
GIN4	0.3350	78.61	Regulates organization of septin ring
BNI5	1.000E-7	13.91	Regulates organization of septin ring
NAP1	1.44E-4	32.55	Regulates organization of septin ring
ELM1	0.9987	96.57	Regulates organization of septin ring
HSL1	0.9920	94.46	Required for degradation of Swe1p
HSL7	0.9997	98.43	Required for degradation of Swe1p
SWE1	0.2575	75.93	Represses cdc28-clb2
MIH1	0.1421	70.69	Activates cdc28-clb2
CDC28	N/A	N/A	Activity required for transition from apical to isotropic bud growth

Supplementary Table 12. TCs defined by knockout of genes thought to be involved in regulation of septin ring formation are listed in the first column. In the second column are tabulated variability p-scores for each TC. The percentile rank (among the total set of 4787 TCs) for the variability p-score is shown in the third column. The fourth column contains a brief description of suspected gene function (see text for details and citations). Several genes were not included in the genetic screen, as their knockouts are lethal; for these TCs, the second and third columns contain an "N/A" designation.

Chapter 3:
Inference of RhoGAP/GTPase Regulation Using Single-cell Morphological Data from a Combinatorial RNAi Screen

Abstract

Biological networks are highly complex, consisting largely of enzymes that act as molecular switches to activate/inhibit downstream targets via post-translational modification. Computational techniques have been developed to perform signaling network inference using some high-throughput data sources, such as those generated from transcriptional and proteomic studies, but no methods have been developed to utilize high-content image-based data, that are emerging principally from large-scale RNAi screens, to these ends. Here, we describe a systematic computational framework for identifying genetic interactions using single-cell morphological data from genetic screens, apply it to GAP/GTPase regulation in *Drosophila*, and evaluate its efficacy. Augmented by knowledge of the basic structure of GAP/GTPase signaling, namely that GAPs act directly upstream of GTPases, we apply our framework for identifying genetic interactions to predict signaling relationships between these proteins. We find that our method makes mediocre predictions using single-knockout morphological data, but achieves vastly improved accuracy by including double-knockout data (sensitivity, 80%, $p < .025$; specificity, 67%). This likely reflects the complex structure of GAP/GTPase signaling, where

each GAP regulates multiple GTPases and each GTPase is regulated by multiple GAPs. We considered other possible methods for inference, and showed that our primary model outperforms the alternatives. Further, we describe a computational framework for identifying genetic interactions; applying this framework to the combinatorial GAP data identifies the biologically validated interaction between RacGAP50C and RacGAP84C. Overall, this work demonstrates the fundamental fact that high-throughput morphological data can be used in a systematic, successful fashion to identify genetic interactions and, using additional knowledge of network structure, to infer signaling relations.

Introduction

Biological signaling networks regulate cellular response to environmental cues. There are still few signaling networks for which a detailed, systems-level description is known, due to the dearth of effective experimental and computational methods [1]. Moreover, these networks are highly complex, consisting largely of enzymes that act as molecular switches to activate/inhibit downstream targets via post-translational modification. These substrates are often themselves enzymes, acting in similar fashion.

Computational techniques have been developed to perform signaling network inference using transcriptional or phosphoproteomic data. These methods typically utilize probabilistic graphical models [2-6] or variations on parameterized modeling [7]. In contrast, using image-based data from genetic screens to predict genetic interactions is significantly more challenging. The range of detectable phenotypes with morphological data is far less than with more traditional data sources: even though dozens or hundreds or geometric morphological features can be defined

and measured on the single-cell level, invariably these features are highly redundant (thus the need for dimensionality reduction). Yet morphological data has the potential to provide information that transcriptional data cannot, namely cellular response to post-translational protein modification.

With the advent of image-based automated technologies and acquisition of high-throughput quantitative imaging data [8, 9], methods have recently been developed which attempt to use these technologies to quantify shape [10], DNA morphology [11], and subcellular-localization of organelles or proteins [12, 13], on a single-cell level. Initial analysis was commonly performed by averaging single-cell results to derive mean scores or by clustering such results [10, 14-16]. Recently, researchers have quantified morphological variability on the single-cell level in response to various stimuli, e.g. genetic or chemical perturbations [17-20]. Classification of cells toward particular phenotypes of interest has been successfully accomplished in multiple cases [8--29]. However, these methods produce (one or more) independent classifiers, each of which is used to score cell similarity to an archetypal shape. Scores from independent classifiers cannot be readily compared one to another, and therefore are a poor framework for systematically scoring putative genetic interactions. This motivated us to produce an alternative framework for classification in which all pairwise relations could be simultaneously scored and compared on equal footing. Indeed, no successful method, to our knowledge, has been developed for systematically predicting genetic interactions or signaling relationships using image-based data from high-throughput screens.

Here we describe a computational framework based on a voting scheme at the single-cell level for identifying genetic interactions utilizing morphological data. We demonstrate the efficacy of this approach by inferring components of the Rho-signaling network in *Drosophila*

melanogaster, namely RhoGAP/GTPase interactions. This network regulates cell adhesion and motility, and perturbations in human orthologs have been implicated in cancer. Rho network structure, with many enzymes and few substrates, is a common network motif [30, 31], and our method makes use of the basic structure of GAP/GTPase signaling, namely that GAPs directly regulate GTPases. To complicate Rho network inference, many predicted *in vitro* enzyme-substrate interactions are not reflected *in vivo* [32].

The core of our method is a classification model that maps putative upstream targets to putative downstream targets on the basis of morphological similarity on the single-cell level following genetic perturbation (RNAi or gene overexpression) of the targets. As input data we utilize a previous image-based screen in the *Drosophila* BG-2 cell line for GTPase overexpression morphological data [10] as well as additional high-throughput combinatorial GAP knockout morphological data published here for the first time (**Supplementary Tables 1**, **2** and **Materials and Methods**). We first apply our method to single-knockout GAP genetic perturbations (also called treatment conditions, or TCs), yielding poor predictions of known GAP/GTPase interactions. Subsequently, by applying our methods to combinatorial double-knockout GAP TCs, we obtain greatly improved predictions of GAP/GTPase interactions. As an additional application of our methodology, we produce an alternative classification model that maps double GAP knockouts to single GAP knockouts, thus providing a means for studying hierarchical relations in GAP regulation. Fundamentally, we show for the first time that high-throughput image-based data can be used with success to predict genetic interactions and, with additional knowledge of network structure, to predict signaling interactions.

Results

We first defined a general classification model (**Fig. 1**) for mapping a set of putative upstream targets (*U*) into a set of putative downstream targets (*D*). We then applied this model to (i) GAP single-knockout TCs (*U*) and GTPase overexpression TCs (*D*), (ii) GAP single- and double-knockout TCs (*U*) and GTPase overexpression TCs (*D*), and (iii) GAP double-knockout TCs (*U*) and GAP single-knockout TCs (*D*).

Classification model for identification of genetic interactions and signaling relationships using morphological data

For the general model, let UTC denote an upstream TC consisting of *c* single cells. The data for UTC consists of a matrix with *c* rows and a column for each morphological feature (in reduced-dimensional feature space; see **Materials and Methods**). To map UTC to one of the elements of *D*, we first classified each single cell in UTC by computing its Mahalanobis distance to each element of *D* and assigning it to the closest downstream TC. The classification of all single cells in UTC may thus be represented by a vector of length *c*, which we termed the *classification vector*. The classification of the cell population, UTC, was defined to be the mode of the classification vector. We calculated a p value for this classification by calculating the probability of observing a mode frequency no smaller than that observed for UTC, using bootstrapping (**Materials and Methods**).

We required that the classification should map each downstream TC to itself with high confidence (i.e., the downstream TCs must be distinguishable from one another); this was true for GTPase overexpression TCs ((i) and (ii)), but not for GAP single-knockouts as the set of

downstream targets (iii). Therefore, a clustering algorithm was developed and implemented as a preprocessing step for the classification model. Following clustering, the classification model successfully mapped each downstream TC to the cluster containing it (**Materials and Methods**).

Double-knockouts are essential for meaningful prediction of signaling relationships using high-throughput morphological data

We first applied our method to map single-knockout GAP TCs to GTPase overexpression TCs (i). We tested the efficacy of our predictions using biologically validated GAP/GTPase interactions from the genes in our dataset (**Supplementary Table 3A**) [33-38] as well as biologically-validated non-interactions (**Supplementary Table 3B**) {cite}. Using single-knockout GAP TCs yielded poor predictions, achieving sensitivity of 2/5 (40%) and specificity of 2/3 (67%) with optimal significance threshold (**Fig. 2A** and **Supplementary Table 4**). We next applied our classification model to map the full set of single- and double-knockout GAP TCs to GTPase overexpression TCs (ii) (**Materials and Methods**). Using the same validation set, we observed vast improvement: the model correctly predicted 4/5 known interactions and 2/3 known non-interactions for an overall sensitivity and specificity of 80% and 67%, respectively (**Fig. 2B**, **2C**, and **Supplementary Table 5**). The method made a total of 12 predictions (out of the 39 possible interactions); the probability of correctly predicting 4/5 known interactions, as determined by hypergeometric statistics, is $p < .025$. This highlights the predictive power of our model as well as the importance of using double-knockout morphological data (**Fig. 2D**, **Supplementary Figs. 1-3**).

Systematic discovery of genetic interactions

We produced an alternative classification model that mapped double GAP knockouts to single GAP knockouts (iii). Here, we treated the set of single-knockouts as the "downstream" targets (*D*). Following clustering to ensure that the classifier mapped each downstream target to the cluster containing it (**Supplementary Table 6**), we applied the model to classify double-knockout GAP TCs to (clusters of) single GAP knockouts (**Supplementary Table 7**). Using this, we constructed a graphical representation of hierarchical relations between pairs of GAPs (**Fig. 3**) by identifying cases of double knockouts, of genes "A" and "B", which were mapped with significance to single-knockout of gene "A". We interpreted this situation as suggestive that protein *A* is required for activity of protein *B*. In effect, this application of the classification model amounts to a systematic way of identifying genetic interactions. Our methods identified the previously validated interaction between RacGAP50C and RacGAP84C (see **Discussion**).

Comparison with alternate methods

This work is the first report of successful signaling inference based on high-throughput signaling. Thus, we considered several alternate methods that might be used to perform inference, and compared these methods to the main classification model developed here.

Mean scores and clustering-based approaches

We calculated mean scores in PC-coordinates in three dimensions, and computed distances from each of the double-knockout TCs to each of the GTPase overexpression TCs. To determine a p-

score for each upstream/downstream pair, we selected samples of equal size to the UTC from the entire set of single cells (for all double-knockout TCs), and computed the distribution of the distance of their mean from the GTPase mean. Applied to the double-knockout GAP data, the mean-score method made many more predictions than the primary classification model presented in the main text. Indeed, in order for the mean-score method to identify 4/5 biologically-validated interactions, it made a total of 23 predictions as compared to 12 for the main classification model, yielding a significance score of only $p = .30$ (as compared to $p < .025$). This highlights one of the caveats of using average morphological data; there is significant variation at the single-cell level within individual TCs [17, 18], making it possible that two TCs' mean scores may resemble each other when their single-cell point clusters do not, thus greatly decreasing the predictive power of a mean-score approach. Interestingly, the mean-score approach correctly classifies the 3 non-interactions, but because of its low predictive power, the single-cell classification model was preferred.

It should be noted that using mean scores is isomorpic to certain clustering-based approaches. One can imagine constructing a classification scheme by defining cutoffs for linkage distances in a hierarchical clustering, for instance. Such an approach is equivalent to computing pairwise distances between mean TC feature scores and identifying the closest pairs. Clustering carries an added disadvantage, namely the possibility of conflating the classification of upstream TCs to downstream TCs; in clustering, all pairwise distances are considered and may influence the final clustering, whereas only pairwise distances between an upstream and downstream TC factor into the mean score method described above (and in our classification model).

Incorporating other classifiers

Neural network classifiers for RacF28L and RhoF30L were previously constructed to classify cells according to similarity with these TCs [10]. We used Z-scores for these two classifiers to represent morphology of each single cell (see **Materials and Methods**), computed mean classifier scores for each double-knockout TC, and ranked TCs accordingly. Using an extremely strict significance cutoff (Bonferroni-corrected $p = .05$), the RacF28L neural networks identified 4 targets (these were a subset of the GAPs predicted by our classification model; namely, RacGAP50C, RacGAP84C, RhoGAP54D, and RhoGAP71E); however, the RhoF30L neural network provided poor specificity, predicting that all 13 GAPs interact with Rho1. The concordance of our results with the predictions of the RacF28L neural network provides added confidence for our findings, but overall this alternative method lacks necessary subtlety to discern genetic interactions more generally, compared with our primary classification model.

Next, again using the two neural network classifiers to represent single-cell morphology, we applied our classification model directly (with just Rac1 and Rho1 as potential downstream targets); this was equivalent to performing a dimensionality reduction using these neural network classifiers, rather than principal components. At optimal threshold, this method attains 60% sensitivity and 67% specificity (**Supplementary Fig. 4**). Again, this performance is poorer than that achieved by our classification model applied to PC-based data.

Discussion

The significance of this work are fourfold. The first contribution is to show the fact that high-throughput morphological data can be used in a systematic fashion identify genetic interactions. Second, we showed the fundamental fact that with additional prior knowledge for the network structure, our framework can be used to identify signaling interactions successfully. Third, the computational framework presented here represents an initial approach to the problem that will serve as a basis for future enhancements (see below). Fourth, and perhaps most intriguing, we showed that our classification model performs much better with both single- and double-knockout data versus only single-knockout data.

In short, we developed a general computational method to predict regulatory interactions using high-throughput image-based data from a genetic screen, and applied it to the case of RhoGAP/GTPase regulation. The method requires some prerequisite knowledge of the structure of GAP/GTPase regulation. Namely, one uses the existence of sequence signatures for GTP-hydrolyzing domains as a means of identification for putative GAPs. Generally speaking, the method proposed here requires additional knowledge of the regulatory structure of the genes under consideration (upstream versus downstream targets). Further development of an unbiased framework for predicting signaling interactions on the sole basis of image-based data is unlikely to be successful due to the high degree of noise present in morphological data and due to the weak informative signal present. Our work here suggests that predictions can be successfully performed using image-based data when combined with additional knowledge, thus could potentially be used to augment predictions using other data sources (e.g., transcriptional) that provide orthogonal information for improved inference.

Why are predictions based on double-knockout TCs better than those using single-knockout TCs? In fact, each GAP likely regulates multiple GTPases and each GTPase is likely regulated

by multiple GAPs. This means that a knockout of a single GAP may not robustly increase activity of any one of the GTPases it normally regulates. However, knockout of two GAPs, each normally regulating the same GTPase, more likely results in increase activity of that GTPase. Put simply, because the regulatory structure is redundant, combinatorial knockouts are necessary to generate a sufficiently informative signal for successful prediction. Our finding in the context of morphological data parallels that of phosphoproteomics data, for which the power of utilizing double-knockouts has been demonstrated [6]. Future work will involve the application of our methods to image-based data related to pathways that are less redundant, for example VEGF (PVR) and MAPK pathways [39, 40].

As an additional application of our methodology, we developed an alternative classification model that maps double GAP knockouts to single GAP knockouts. Viewed generally, this methodology actually represents a systematic way to identify genetic interactions using quantitative morphological data [41]. Applied to the specific case of GAPs, the methodology provides a means for probing hierarchical relations between pairs of GAPs. Of particular interest is the interaction between two GAPs regulating the *same* GTPase. A dosage response interaction has been described between RacGAP50C and RacGAP84C in fly wing [35]. Here we found RacGAP50C-/RacGAP84+ and RacGAP50C-/RacGAP84- TCs share significant morphological similarity at the single-cell level, suggesting that RacGAP50C is required for RacGAP84C activity. Since RacGAP50C and RacGAP84C both signal through Rac1, and the fact that Rac1 is likely a "date hub" [42], one possible explanation for this set of observations is that RacGAP50C and RacGAP84C interact with Rac1 in a process-dependent manner, with the RacGAP50C/Rac1 interaction occurring earlier than that of RacGAP84/Rac1.

A potential objection to our method of validation is the relative dearth of positive control data. On the contrary, we propose that our model's predictions for novel GAP/GTPase interactions could serve as targets for further study by biological means. Human and yeast data suggest that many more GAP/GTPase interactions likely occur in fly than have been previously validated [43], meaning that we should expect the classification model to generate a large number of false positives, which correspond to interactions that have not been previously validated biologically and were thus not included in our validation set.

Future work will involve acquisition of additional double-knockout morphological data to allow for prediction of other known GTPase targets, as well as for better simultaneous predictions of multiple GTPase targets for a single GAP knockout. For the latter task, one possibility would be to obtain double-overexpression GTPase data and augment the classification model with these TCs as targets. A GAP treatment condition mapped to a double-overexpression class (versus either of the single overexpression classes) would suggest multiple GTPase targets for the GAP. Additional work will involve application of our methods to new image-based data sources, as well as integration with methods utilizing transcriptional data for improved inference of signaling networks.

Materials and Methods

Morphological datasets

GTPase overexpression

As described in [10], TCs were prepared in the *Drosophila* DM-BG2 (referred to as BG-2) cell line using either dsRNA or overexpression constructs. The screen consisted of 249 distinct genetic perturbations, with several replicates, for a total of 273 TCs, including two treatment conditions corresponding to constitutively active Rac1 (RacF28L) and Rho1 (RhoF30L) mutants, respectively, and a treatment condition corresponding to a fast-cycling Cdc42 mutant (Cdc42Y32A). For each single cell in each treatment condition, 145 geometric features and 9 status features (**Supplementary Table 1**) were extracted in a semi-automated fashion. In total, 12601 single cells were imaged, for an average of 46 single cells for each TC.

GAP single- and double-knockouts

Drosophila BG-2 cells were transfected with dsRNAs targeting 13 RhoGAPs (**Supplementary Table 2**) in all possible combination components in combination with **act-GAL4** and **UAS-GFP** plasmids. Live cells were imaged and the morphology of single cells was quantified using previously described methods. Cell segmentation was performed using the custom CellSegmenter Software. Stochastic labeling with GFP was used to facilitate image segmentation. For each single cell, the same 145 geometric and 9 status features were extracted. All 13 single-knockout TCs were constructed and all except one (RhoGAP19D/RhoGAP54D) of the $\binom{13}{2} = 78$ possible double-knockout TCs were successfully constructed, for a total of 90 TCs. Overall, 6480 single cells were imaged, for an average of 72 cells per TC.

Data normalization and dimensionality reduction

Normalization and dimensionality reduction was performed for the 273-TC dataset [17]. Briefly, each of the 145 raw features was normalized to have mean 0 and variance 1 across the full set of 12601 single cells. Normalization of the raw features was done to avoid inappropriately weighting some features over others (for example, because of arbitrary differences in unit measurements). Following normalization, dimensionality reduction was performed by computing principal components (PCs) for the full set of single-cell data, and then projecting each data point onto the first three PCs. Working in reduced feature space avoided inappropriately weighting particular morphological feature classes that are overrepresented in the set of raw features (for example, redundant measurements of nucleus shape).

Similarly, normalization was performed for the 90-TC dataset. Dimensionality reduction was performed using the first three PCs computed using the 273-TC dataset. We used principal components from the 273-TC dataset in order to readily compare GAP knockouts and GTPase overexpression TCs. We chose to use the larger dataset because it contained knockout, overexpression, and control test data, and because previous work [17] had shown robustness in dimensionality reduction. We repeated similar robustness testing here, finding that our subsequent analysis was robust to varying the number of dimensions of reduced feature space (**Supplementary Table 8**).

Classification model

While clustering elicits some structural features of the data, our primary goal was to develop a classification model mapping the set, U, of upstream (GAP knockout) TCs into the set, D, of downstream (GTPase overexpression) TCs. It was desirable that our model should (1) utilize

single-cell data, rather than mean scores for each TC, (2) assign meaningful confidence scores to each classification, and (3) correctly classify control (GTPase overexpression) TCs.

Let $U = \{UTC_1, UTC_2, \ldots, UTC_n\}$ and $D = \{DTC_1, DTC_2, \ldots, DTC_m\}$, where UTC_i denotes the i^{th} upstream TC and DTC_j denotes the j^{th} downstream TC. Let c_i denote the number of single cells in UTC_i. To classify UTC_i into D, first each of its c_i single cells is separately classified into D by calculating the Mahalanobis distance to each DTC_j and selecting the closest DTC_j. The classification of single cells in UTC_i can thus be represented as a vector of length c_i, which we call the *classification vector*. The classification of UTC_i, denoted $f(UTC_i)$, is defined to be the mode of the classification vector. Let d_i denote the frequency of the mode (note that $\frac{c_i}{m} \leq d_i \leq c_i$). In other words, we map UTC_i to the DTC_j onto which the greatest number (d_i) of single cells in UTC_i were mapped. Intuitively, the classification of UTC_iincreases in confidence as $d_i \rightarrow c_i$. In case of a tie for the mode of the classification vector, UTC_i is mapped to two (or more) downstream targets. Ties are rare when the TC size c_i is large. For GAP single- and double-knockout data mapped into GTPase overexpression, four instances of ties occurred, but none of these classifications had statistically significant confidences.

Confidence scores were assigned using bootstrapping to make rigorous the intuition that classification increases in confidence as $d_i \rightarrow c_i$. Specifically, the confidence of the classification of UTC_i, consisting of c_i single cells, was determined by selecting 1000 random samples of c_i cells taken from the full set of upstream TCs (i.e., the full set of 6480 single cells), classifying these samples into D using the above method, and calculating the distribution of the mode frequency, d, of the classification vector across the set of bootstrapped samples. This

distribution was used to determine the probability of observing a classification vector mode frequency no smaller than that observed for the classification of UTC_i, i.e. the probability that $d \geq d_i$.

An additional requirement of the model was that it should correctly classify downstream targets onto themselves with high confidence. To verify this, we treated each DTC_j temporarily as a member of U and applied the classification algorithm, thus mapping each DTC_j into D. Mathematically, the condition we required was simply that $f: D \rightarrow D$ is a bijection. Intuitively, this tested whether the $\{DTC_j\}$ are distinguishable in reduced feature space. If the single-cell clusters for two downstream TCs overlap, then the one of these TCs may be mapped into the other, or mapped into itself with low confidence.

We applied this general framework to classify the set of GAP single- and double-knockout TCs (U) into the set of GTPase overexpression TCs (D) (**Fig. 2**, **Supplementary Fig. 1**). For single-knockout data, results were not significantly by drawing samples from the set of cells comprising only single-knockout TCs versus the entire set of single- and double-knockouts (**Supplementary Table 9**). As required, the model correctly classifies each GTPase overexpression experiment with high confidence (**Supplementary Table 10**). For double-knockout GAP TCs, e.g. knockout of GAPs "A" and "B", we interpreted a positive classification to GTPase "C" to suggest that both A and B signal through C, unless the single-knockout GAP TC for either "A" or "B" was classified to "C" at Bonferroni-corrected $p = .05$ (in which case the double-knockout "A" and "B" was not considered). This reflects the structure of GAP/GTPase signaling, where multiple GAPs regulate the same GTPase, meaning that multiple knockouts of GAPs may be necessary to observe increased activity of the GTPase they co-regulate. We exclude

consideration of a double-knockout in the case that one of the single knockout components was classified at high significance to the same GTPase, as this was necessary to avoid false positive predictions associated with single-knockouts that dominate morphology (in practice, this excludes double-knockouts with RhoGAP92B for the primary classification model). We also incorporated this exclusion into the alternative algorithms (mean-score methods, neural network-based methods) under consideration so that the algorithms could be judged fairly. We verified that the classification model is robust to noise in input data, particularly for TCs that were classified with high confidence, by jackknife statistics (**Supplementary Fig. 5** and **Supplementary Table 11**).

Mapping double-knockouts into single-knockouts

As an additional application of our classification model, we applied it to the set of GAP double-knockouts (U) and GAP single-knockouts (D). Applying the model directly to the entire set, D, was not possible, because each element of D was not correctly mapped to itself. That is, some single-knockout TCs were classified into different single-knockout TCs, due to the fact that some of 13 single-knockout TCs were not morphologically distinguishable from one another. To remedy this, we clustered the single-knockout TCs using a variant of EM designed to guarantee that, under the final clustering, all single-knockout TCs would be correctly classified (**Supplementary Fig. 6**). The algorithm proceeds by iterating the following two steps, beginning with $k = 0$ and $D_0 = D$.

Iteration k:

(i) Map each element of D into D_k using the classification. If each element of D is mapped to the cluster containing it, set $\tilde{D} = D_k$ and exit.

(ii) Define D_{k+1} as follows. Let $f(D) = R_k = \{r_i\}$ denote the range of the classification of D mapped into D_k (i.e. the subset of D_k onto which elements of D were mapped in (i)). Then set

$$D_{k+1} = \{\cup f^{-1}(r_i)\}.$$

At each iteration, the elements of D mapped to the same target in D_k are grouped (by taking the union of single cells comprising each such element) into a single element of D_{k+1}. Note that upon termination of the algorithm, the clustering $\tilde{D}$ necessarily has the property that every element of D is classified to the cluster containing it. It is theoretically possible for the algorithm to enter a cycle (though unlikely; this did not occur for our test data), in which case all elements forming a cycle are clustered together, thus allowing the algorithm to continue. In the worst case, the algorithm terminates by grouping all elements of D into a single element, which has to be mapped to itself.

For single-knockout GAP TCs, the clustering algorithm terminates with $\tilde{D} = D_2$, yielding a total of 5 clusters (**Supplementary Table 6**). (By comparison, for GTPase ovexpression TCs, the clustering algorithm terminates immediately, i.e. $\tilde{D} = D_0$.) The classification model was used to map all double-knockout GAP TCs into the set $\tilde{D}$ of clustered single-knockout GAP TCs (**Fig. 3** and **Supplementary Table 7**).

References

1. Friedman, A., Perrimon, N.: Genetic screening for signal transduction in the era of network biology. Cell 128 (2007) 225–231

2. Friedman, N.: Inferring cellular networks using probabilistic graphical models. Science 303(5659) (2004) 799–805

3. Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D.A., Nolan, G.P.: Causal protein-signaling networks derived from multiparameter single-cell data. Science 308(5721) (2005) 523–529

4. Friedman, N., Linial, M., Nachman, I., Pe'er, D.: Using Bayesian networks to analyze expression data. J. of Computational Biology 7(3-4) (2000) 601–620

5. Peer, D., Regev, A., Elidan, G., Friedman, N.: Inferring subnetworks from perturbed expression profiles. Bioinformatics 17 (2001) S214–S224

6. Bakal C, Linding R, Llense F, Heffern E, Martin-Blanco E, Pawson T, Perrimon N. Phosphorylation Networks Regulating JNK Activity in Diverse Genetic Backgrounds. Science. 2008 Oct 17;322(5900):453-456.

7. Baym M, Bakal C, Perrimon N, Berger B.: High-Resolution Modeling of Cellular Signaling Networks. Proceedings of the 12th Annual International Conference on Research in Computational Molecular Biology (RECOMB 2008), LNBI 4955: 257-271, 2008

8. Carpenter AE, Jones TR, Lamprecht MR, Clarke C, Kang IH, Friman O, Guertin DA, Chang JH, Lindquist RA, Moffat J, Golland P, Sabatini DM. CellProfiler: image analysis software for identifying and quantifying cell phenotypes. Genome Biol. 2006 ;7(10):R100.

9. Ohya Y, Sese J, Yukawa M, Sano F, Nakatani Y, Saito TL, Saka A, Fukuda T, Ishihara S, Oka S, Suzuki G, Watanabe M, Hirata A, Ohtani M, Sawai H, Fraysse N, Latgé J, François JM, Aebi M, Tanaka S, Muramatsu S, Araki H, Sonoike K, Nogami S, Morishita S. High-dimensional and large-scale phenotyping of yeast mutants. Proc Natl Acad Sci U S A. 2005 Dec 27;102(52):19015-20.

10. Bakal C, Aach J, Church G, Perrimon N. Quantitative Morphological Signatures Define Local Signaling Networks Regulating Cell Morphology. Science. 2007 Jun 22;316(5832):1753-1756.

11. Moffat J, Grueneberg DA, Yang X, Kim SY, Kloepfer AM, Hinkle G, Piqani B, Eisenhaure TM, Luo B, Grenier JK, Carpenter AE, Foo SY, Stewart SA, Stockwell BR, Hacohen N, Hahn WC, Lander ES, Sabatini DM, Root DE. A lentiviral RNAi library for human and mouse genes applied to an arrayed viral high-content screen. Cell. 2006 Mar 24;124(6):1283-98.

12. Glory E, Murphy RF. Automated subcellular location determination and high-throughput microscopy. Dev Cell. 2007 Jan ;12(1):7-16.

13. Perlman ZE, Slack MD, Feng Y, Mitchison TJ, Wu LF, Altschuler SJ. Multidimensional drug profiling by automated microscopy. Science. 2004 Nov 12;306(5699):1194-8.

14. Neumann B, Held M, Liebel U, Erfle H, Rogers P, Pepperkok R, Ellenberg J. High-throughput RNAi screening by time-lapse imaging of live human cells. Nat Methods. 2006 May ;3(5):385-90.

15. Piano F, Schetter AJ, Morton DG, Gunsalus KC, Reinke V, Kim SK, Kemphues KJ. Gene clustering based on RNAi phenotypes of ovary-enriched genes in C. elegans. Curr Biol. 2002 Nov 19;12(22):1959-64.

16. Gil J, Wu H, Wang BY. Image analysis and morphometry in the diagnosis of breast cancer. Microsc Res Tech. 2002 Oct 15;59(2):109-18.

17. Nir et al Variability

18. Levy SF, Siegal ML. Network Hubs Buffer Environmental Variation in Saccharomyces cerevisiae. PLoS Biol. 2008 Nov 4;6(11):e264.

19. Slack MD, Martinez ED, Wu LF, Altschuler SJ. Characterizing heterogeneous cellular responses to perturbations. Proc. Natl. Acad. Sci. U.S.A. 2008 Dec 9;105(49):19306-19311.

20. Spencer SL, Gaudet S, Albeck JG, Burke JM, Sorger PK. Non-genetic origins of cell-to-cell variability in TRAIL-induced apoptosis. Nature. 2009 May 21;459(7245):428-432.

21. Jones TR, Carpenter AE, Lamprecht MR, Moffat J, Silver SJ, Grenier JK, Castoreno AB, Eggert US, Root DE, Golland P, Sabatini DM. Scoring diverse cellular morphologies in image-based screens with iterative feedback and machine learning. Proceedings of the National Academy of Sciences. 2009 Feb 10;106(6):1826-1831.

22. Boland MV, Murphy RF. A neural network classifier capable of recognizing the patterns of all major subcellular structures in fluorescence microscope images of HeLa cells. Bioinformatics. 2001 Dec 1;17(12):1213-1223.

23. Tanaka M, Bateman R, Rauh D, Vaisberg E, Ramachandani S, Zhang C, Hansen KC, Burlingame AL, Trautman JK, Shokat KM, Adams CL. An Unbiased Cell Morphology–Based Screen for New, Biologically Active Small Molecules. PLoS Biol. 2005 Apr 5;3(5):e128.

24. Chen X, Murphy RF. Automated interpretation of protein subcellular location patterns. Int. Rev. Cytol. 2006 ;249193-227.

25. Boland MV, Mia K. Markey, Robert F. Murphy. Automated recognition of patterns characteristic of subcellular structures in fluorescence microscopy images. Cytometry. 1998 ;33(3):366-375.

26. Wang J, Zhou X, Bradley PL, Chang S, Perrimon N, Wong ST. Cellular Phenotype Recognition for High-Content RNA Interference Genome-Wide Screening. J Biomol Screen. 2008 Jan 1;13(1):29-39.

27. Adams CL, Kutsyy V, Coleman DA, Cong G, Crompton AM, Elias KA, Oestreicher DR, Trautman JK, Vaisberg E. Compound classification using image-based cellular phenotypes. Meth. Enzymol. 2006 ;414440-468.

28. Loo L, Wu LF, Altschuler SJ. Image-based multivariate profiling of drug responses from single cells. Nat. Methods. 2007 May ;4(5):445-453.

29. Young DW, Bender A, Hoyt J, McWhinnie E, Chirn G, Tao CY, Tallarico JA, Labow M, Jenkins JL, Mitchison TJ, Feng Y. Integrating high-content screening and ligand-target prediction to identify mechanism of action. Nat Chem Biol. 2008 Jan ;4(1):59-68.

30. Albert, R.: Scale-free networks in cell biology. J Cell Sci 118(21) (2005) 4947–4957

31. Csete, M., Doyle, J.: Bow ties, metabolism and disease. Trends in Biotechnology 22(9) (2004) 446–450

32. Michiels, F., Habets, G.G.M., Stam, J.C., van der Kammen, R.A., Collard, J.G.: A role for rac in tiaml-induced membrane ruffling and invasion. Nature 375 (1995) 338–340

33. G. Grumbling, V. Strelets and The FlyBase Consortium (2006). FlyBase: anatomical data, images and queries. *Nucleic Acids Research* **34:** D484-D488; doi:10.1093/nar/gkj068. http://flybase.org/

34. Stark C, Breitkreutz B, Reguly T, Boucher L, Breitkreutz A, Tyers M. BioGRID: a general repository for interaction datasets. Nucleic Acids Res. 2006 Jan 1;34(Database issue):D535-539.

35. Sotillos S, Campuzano S. DRacGAP, a novel Drosophila gene, inhibits EGFR/Ras signalling in the developing imaginal wing disc. Development. 2000 Dec 15;127(24):5427-5438.

36. Billuart P, Winter CG, Maresh A, Zhao X, Luo L. Regulating axon branch stability: the role of p190 RhoGAP in repressing a retraction signaling pathway. Cell. 2001 Oct 19;107(2):195-207.

37. Raymond K, Bergeret E, Dagher M, Breton R, Griffin-Shea R, Fauvarque M. The Rac GTPase-activating Protein RotundRacGAP Interferes with Drac1 and Dcdc42 Signalling in Drosophila melanogaster. J. Biol. Chem. 2001 Sep 14;276(38):35909-35916.

38. Lundström A, Gallio M, Englund C, Steneberg P, Hemphälä J, Aspenström P, Keleman K, Falileeva L, Dickson BJ, Samakovlis C. Vilse, a conserved Rac/Cdc42 GAP mediating Robo repulsion in tracheal cells and axons. Genes Dev. 2004 Sep 1;18(17):2161-2171.

39. Kiger A, Baum B, Jones S, Jones M, Coulson A, Echeverri C, Perrimon N. A functional genomic analysis of cell morphology using RNA interference. Journal of Biology. 2003 ;2(4):27.

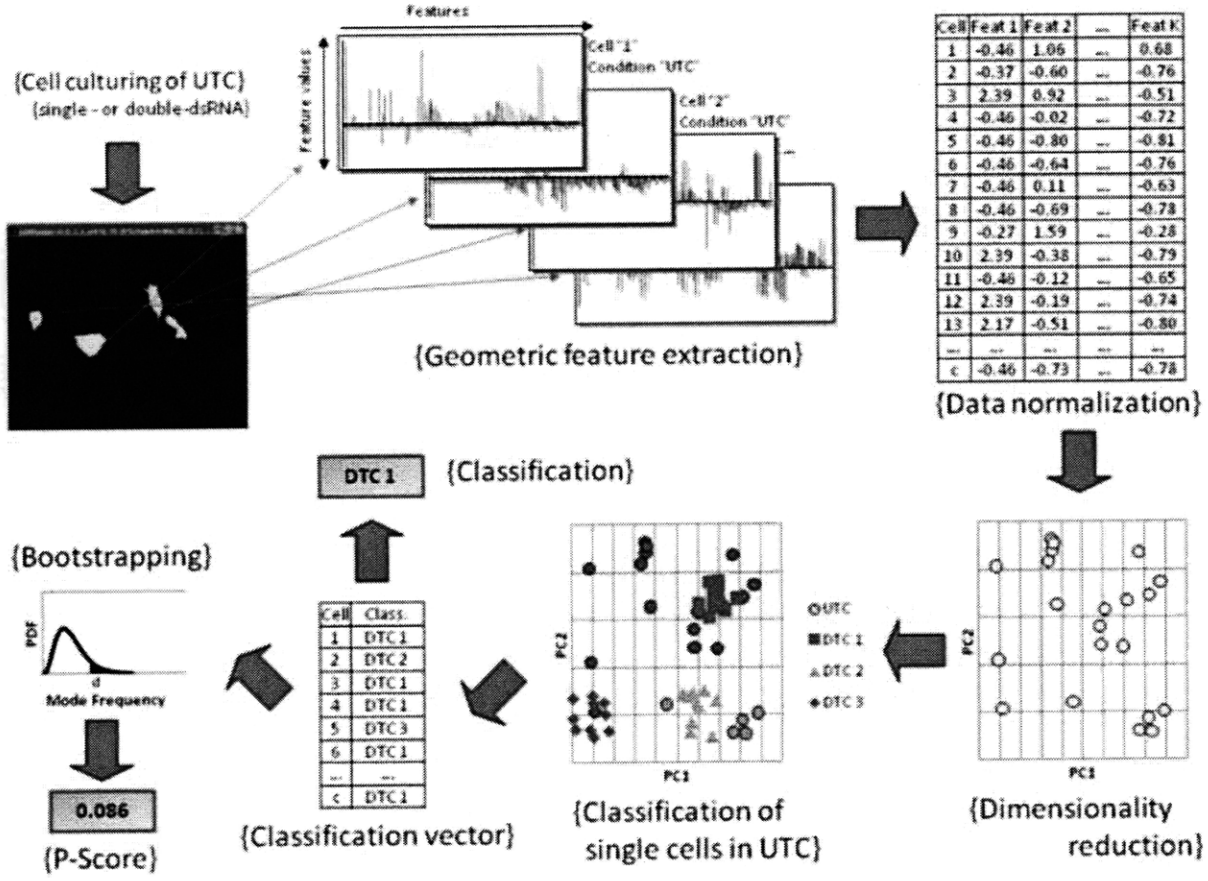
40. Sims D, Duchek P, Baum B. PDGF/VEGF signaling controls cell size in Drosophila. Genome Biol. 2009 Feb 12;10(2):R20.

41. Tong AHY, Lesage G, Bader GD, Ding H, Xu H, Xin X, Young J, Berriz GF, Brost RL, Chang M, Chen Y, Cheng X, Chua G, Friesen H, Goldberg DS, Haynes J, Humphries C, He G, Hussein S, Ke L, Krogan N, Li Z, Levinson JN, Lu H, Ménard P, Munyana C, Parsons AB, Ryan O, Tonikian R, Roberts T, Sdicu A, Shapiro J, Sheikh B, Suter B, Wong SL, Zhang LV, Zhu H, Burd CG, Munro S, Sander C, Rine J, Greenblatt J, Peter M, Bretscher A, Bell G, Roth FP, Brown GW, Andrews B, Bussey H, Boone C. Global mapping of the yeast genetic interaction network. Science. 2004 Feb 6;303(5659):808-813.

42. Han JJ, Bertin N, Hao T, Goldberg DS, Berriz GF, Zhang LV, Dupuy D, Walhout AJM, Cusick ME, Roth FP, Vidal M. Evidence for dynamically organized modularity in the yeast protein-protein interaction network. Nature. 2004 Jul 1;430(6995):88-93.

43. Yu J, Pacifico S, Liu G, Finley RL. DroID: the Drosophila Interactions Database, a comprehensive resource for annotated gene and protein interactions. BMC Genomics. 2008 ;9461.

Figure 1: Workflow for classification of upstream targets (e.g., RhoGAPs) to downstream targets (e.g., RhoGTPases) using high-throughput morphological data.



Cell	Feat 1	Feat 2	...	Feat K
1	-0.46	1.06	...	0.68
2	-0.37	-0.60	...	-0.76
3	2.39	0.92	...	-0.51
4	-0.46	-0.02	...	-0.72
5	-0.46	-0.80	...	-0.81
6	-0.46	-0.64	...	-0.76
7	-0.46	0.11	...	-0.63
8	-0.46	-0.69	...	-0.78
9	-0.27	1.59	...	-0.28
10	2.39	-0.38	...	-0.79
11	-0.46	-0.12	...	-0.65
12	2.39	-0.19	...	-0.74
13	2.17	-0.51	...	-0.80
...	...	...	...	...
c	-0.46	-0.73	...	-0.78

Cell	Class
1	DTC 1
2	DTC 2
3	DTC 1
4	DTC 1
5	DTC 3
6	DTC 1
...	...
c	DTC 1

Cell culture was subjected to a variety of genetic perturbations, multiple single-cell images were acquired for each treatment condition, and raw geometric features were extracted for each single cell (upper left, upper middle). The raw data was subjected to normalization (top right) and dimensionality reduction. The *c* single cells comprising each downstream TC and upstream TC were represented as points in reduced feature space (bottom right, shown for UTC). Given a particular UTC, each of its cells was mapped to one of the DTCs using a classification map based on computing a modified Euclidean distance to each DTC point-cluster and selecting the

closet DTC; single-cell results were compiled in the classification vector (bottom middle). The classification for UTC, in turn, was defined to be the mode of the classification vector. Subsequently, bootstrapping was performed to determine the distribution of the mode frequency for samples of size *c* drawn from the full set of single cells from all UTCs and classified onto the set of DTCs; this distribution was used to calculate the p-score for the classification of UTC (bottom left). See text for additional details.

Figure 2: Inference of RhoGAP/GTPase regulation using morphological data from single- versus double-knockout GAP treatment conditions.

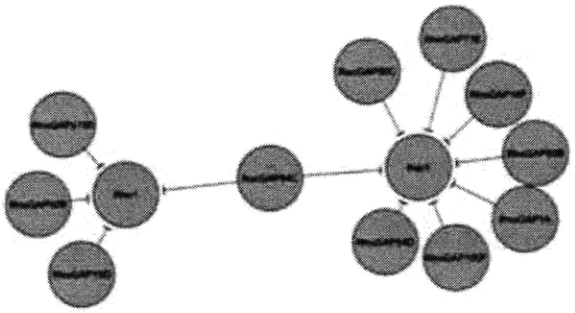
A

GAP	GTPase
RacGAP50C	**Rac1**
RhoGAP54D	Rac1
CdGAPr	Rac1
RhoGAPp190	**Rho1**
RhoGAP19D	Rho1
RhoGAP92B	Rho1
RacGAP84C	*Rho1*

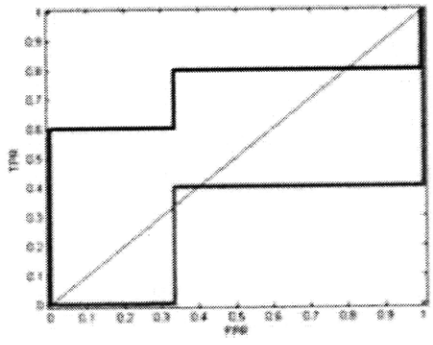
B

GAP	GTPase
RacGAP50C	**Rac1**
RacGAP84C	**Rac1**
RhoGAP93B	**Rac1**
RhoGAP54D	Rac1
RhoGAP71E	**Rac1**
RhoGAP1A	Rac1
RhoGAP16F	**Rac1**
RhoGAP100F	Rac1
RhoGAPp190	**Rho1**
RhoGAP19D	Rho1
RhoGAP92B	**Rho1**
RacGAP84C	*Rho1*

C



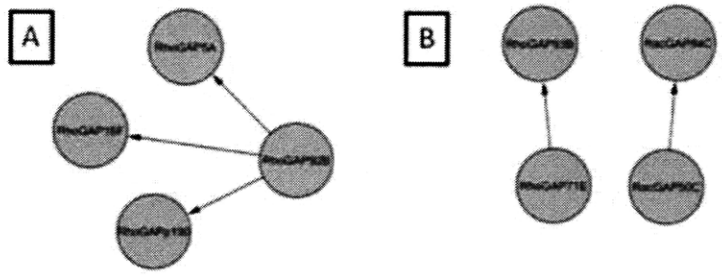
D



(A) Classification of single-knockout GAP TCs to GTPase overexpression TCs. The threshold was selected to yield an optimal model, according to ROC analysis. A total of 7 predictions were made, 2/5 biologically-validated interactions (bolded) were correctly predicted, and 2/3 non-interactions were correctly predicted (the one false positive is italicized). **(B)** Classification of both single- and double-knockout GAP TCs to GTPase overexpression TCs. All pairs listed here are significant at optimal threshold, as determined by ROC analysis. The model correctly predicts 4/5 biologically-validated interactions (bolded) and 2/3 non-interactions (the one false positive is italicized). Overall, the model made 12 out of 39 possible predictions, yielding a p-score of $p < .025$ for identifying 4/5 positive interactions. The model mapped

several GAPs to Cdc42, but none with sufficient significance (see **Supplementary Table 5** for full results). **(C)** Network representation of predicted signaling interactions for the double-knockout classification model. **(D)** ROC curve showing single-knockout (red) versus double-knockout (blue) predictive models. For the single-knockout model, the optimal threshold yields the only model that makes better predictions than random guessing. For the double-knockout model, given that the set of validated interactions is likely incomplete, we err on the side of producing more false positives, and prefer (.33, .80) to (0, .60).

Figure 3: Hierarchical GAP relations demonstrating genetic interactions predicted by the classification model



Hierarchical relations between GAPs based on classification of double-knockout GAP TCs into single-knockout GAP TCs. **(A) Rho1 hierarchy.** Double-knockout TCs for RhoGAP92B/RhoGAP5A, RhoGAP92B/RhoGAP16F, and RhoGAP92B/RhoGAPp190 all shared significant morphological similarity with single-knockout of RhoGAP92B; i.e. these double-knockout TCs were mapped to the cluster containing the RhoGAP92B single-knockout TC, and bootstrapping yielded p-scores for this classification that were significant at $p = .05$ following Bonferroni correction. Furthermore, none of these three proteins was in the same cluster with RhoGAP92B (**Supplementary Table 6**). We previously predicted that all four proteins signal through Rho1. Taken together, these observations suggest that activity of RhoGAP92B may be required for the repressive activity of RhoGAP5A, RhoGAP16F, and RhoGAPp190, respectively, on Rho1. **(B) Rac1 hierarchy.** Analogous results were obtained for RacGAP50C/RacGAP84C: the double-knockout resembled the single-knockout of RacGAP50C, the single-knockout of RacGAP84C was in a different cluster than RacGAP50C, and both of these proteins were previously predicted to signal through Rac1. These observations suggest that RacGAP50C may be required for the activity of RacGAP84C on Rac1. In the final

case, the double-knockout RhoGAP71E/RhoGAP93B resembled the single-knockout of RhoGAP71E, and this single-knockout was clustered distinctly from the single-knockout of RhoGAP93B. In this case, our previous classification mapped RhoGAP71E but not RhoGAP93B to Rac1; however, the interaction between RhoGAP93B and Rac1 has been biologically validated. Taken together, then, these results suggest that RhoGAP71E may be required for the activity of RhoGAP93B on Rac1.

Inference of RhoGAP/GTPase Regulation Using Single-cell Morphological Data from a Combinatorial Genetic Screen

Oaz Nir, Chris Bakal, Norbert Perrimon & Bonnie Berger

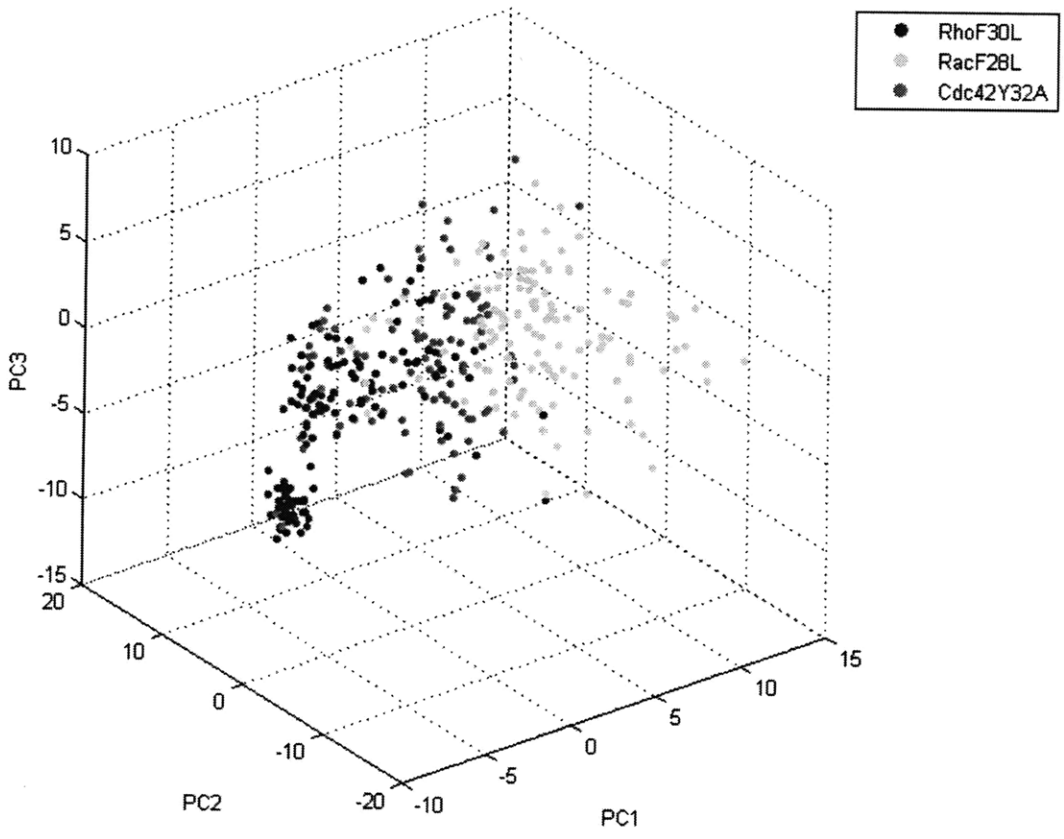
Supplementary figures and text:

Supplementary Figure 1	Point sets for GTPase overexpression TCs and classification of all single cells in the double-knockout screen
Supplementary Figure 2	RacGAP50C and RacGAP84C single- and double-knockouts
Supplementary Figure 3	RacGAP50C and RhoGAP93B single- and double-knockouts
Supplementary Figure 4	ROC curve for neural network based-alternative classification model
Supplementary Figure 5	Robustness of classification to exclusion of data using jackknifing
Supplementary Figure 6	Classification-based clustering algorithm for downstream target TCs
Supplementary Table 1	List of raw geometric features for *Drosophila* screens
Supplementary Table 2	List of GAPs included in genetic screen
Supplementary Table 3	Biologically validated RhoGAP/GTPase interactions and non-interactions
Supplementary Table 4	Classification of single-knockout GAP TCs into GTPase overexpression TCs
Supplementary Table 5	Classification of single- and double-knockout GAP TCs into GTPase overexpression TCs
Supplementary Table 6	Clustering of single-knockout GAP TCs
Supplementary Table 7	Classification of double-knockout GAP TCs into single-knockout GAP TCs
Supplementary Table 8	Robustness of classification to method of dimensionality reduction
Supplementary Table 9	Alternative bootstrapping for mapping single-knockout GAP TCs to GTPase overexpression TCs
Supplementary Table 10	Classification of the set of GTPase overexpression TCs to itself
Supplementary Table 11	Robustness of classification to exclusion of data using jackknifing
Supplementary Table 12	Sensitivity/specificity for mapping single- and double-knockout GAP TCs to GTPase overexpression TCs

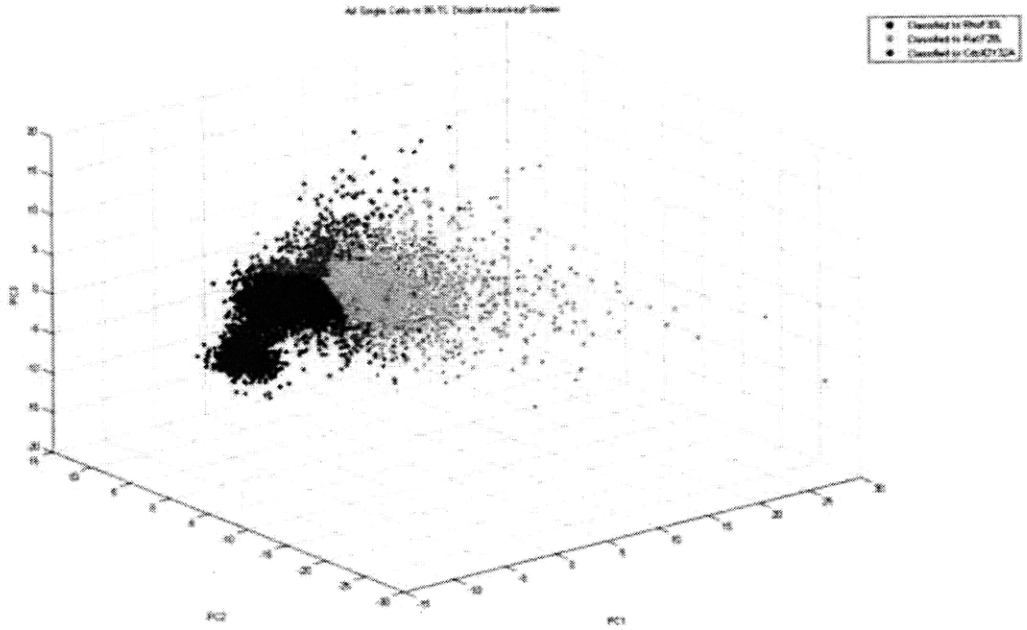
Supplementary Figure 1

Point sets for GTPase overexpression TCs and classification of all single cells in the double-knockout screen

Supplementary Fig. 1A: Point sets for RhoF30L, RacF28L, and Cdc42Y32A



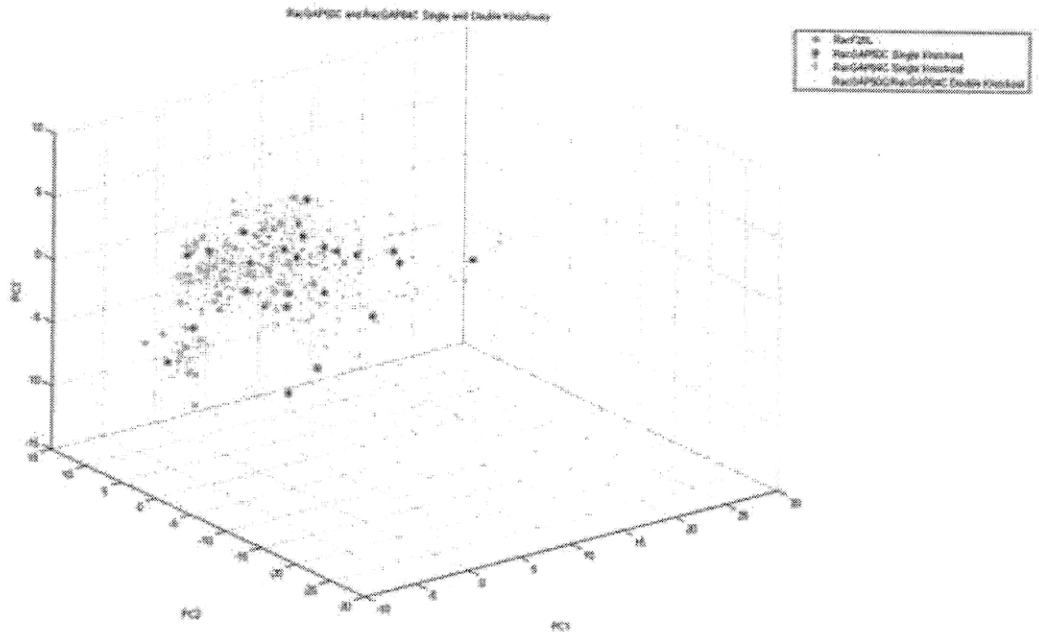
Supplementary Fig. 1B: Mapping of all single cells from the double-knockout GAP screen to GTPase overexpression TCs



Supplementary Fig. 1. Point sets for GTPase overexpression TCs and classification of all single cells in the double-knockout screen. **(A)** Point sets for RhoF30L (blue), RacF28L (green), and Cdc42Y32A (red) shown in reduced-dimensional feature space. **(B)** The mapping of all 6480 single cells from the double-knockout GAP screen to GTPase overexpression TCs. Overall, the classification model defines a phase space for mapping single cells in the set of upstream TCs to the set of downstream TCs.

Supplementary Figure 2

RacGAP50C and RacGAP84C single- and double-knockouts

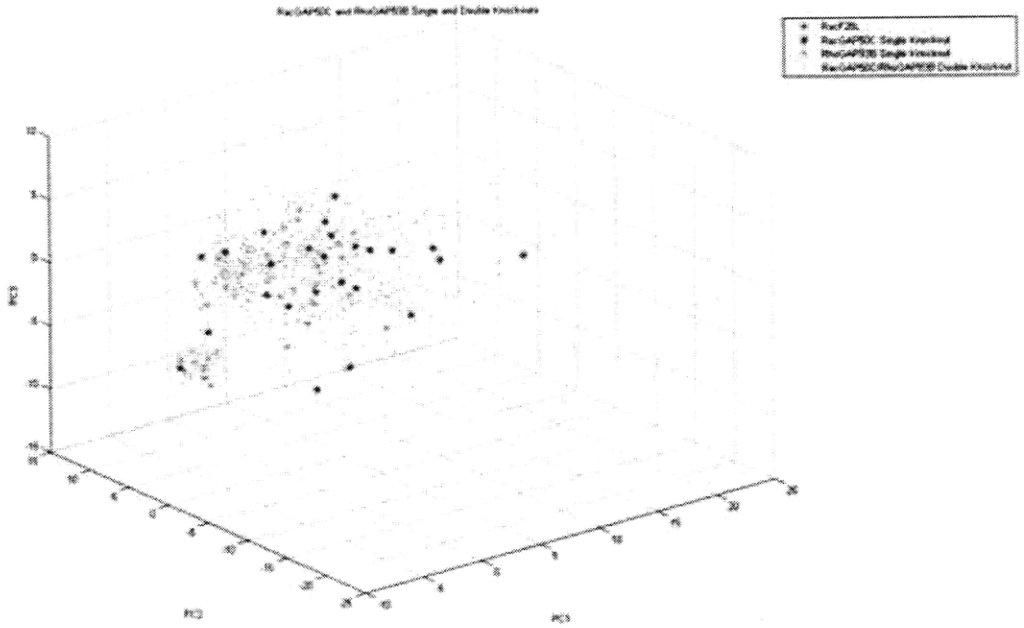


Supplementary Fig. 2. RacGAP50C and RacGAP84C single- and double-knockouts. The plot in PC-based coordinates shows single-cell point sets for RacF28L (green), RacGAP50C single-knockout (magenta), RacGAP84C single-knockout (cyan), and RacGAP50C/RacGAP84C double-knockout (yellow). The classification model maps the RacGAP50C single-knockout to RacF28L with low confidence and actually maps the RacGAP84C single-knockout to Rho1 with high confidence (incorrectly), but it maps the RacGAP50C/RacGAP84C double-knockout to Rac1 with high confidence (correctly).

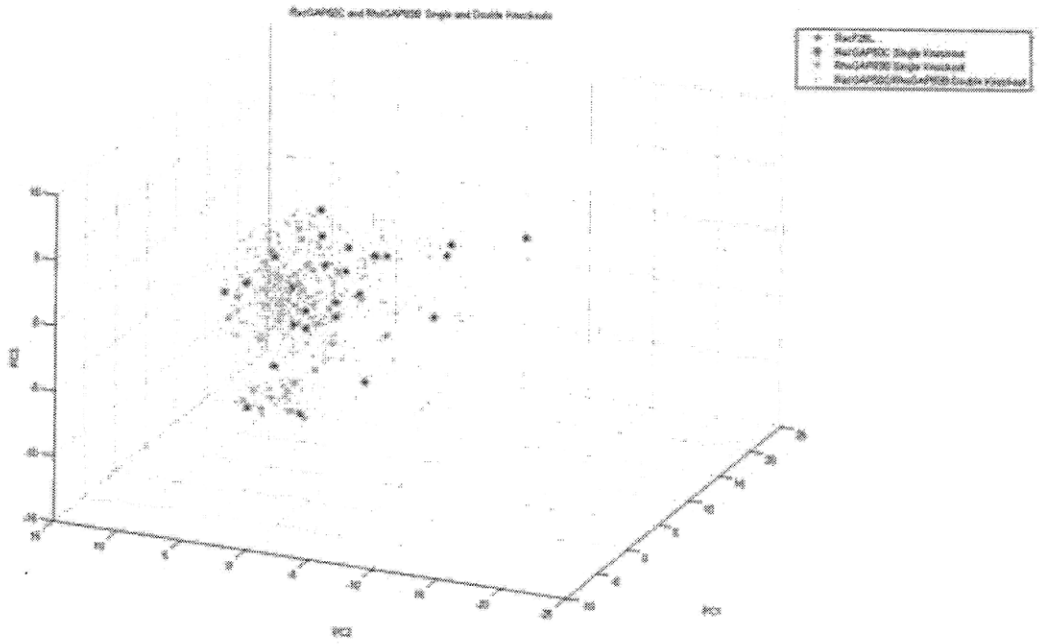
Supplementary Figure 3

RacGAP50C and RhoGAP93B single- and double-knockouts

Supplementary Fig. 3A: Point sets for RacF28L, RacGAP50C single-knockout, RhoGAP93B single-knockout, and RacGAP50C/RhoGAP93B double-knockout



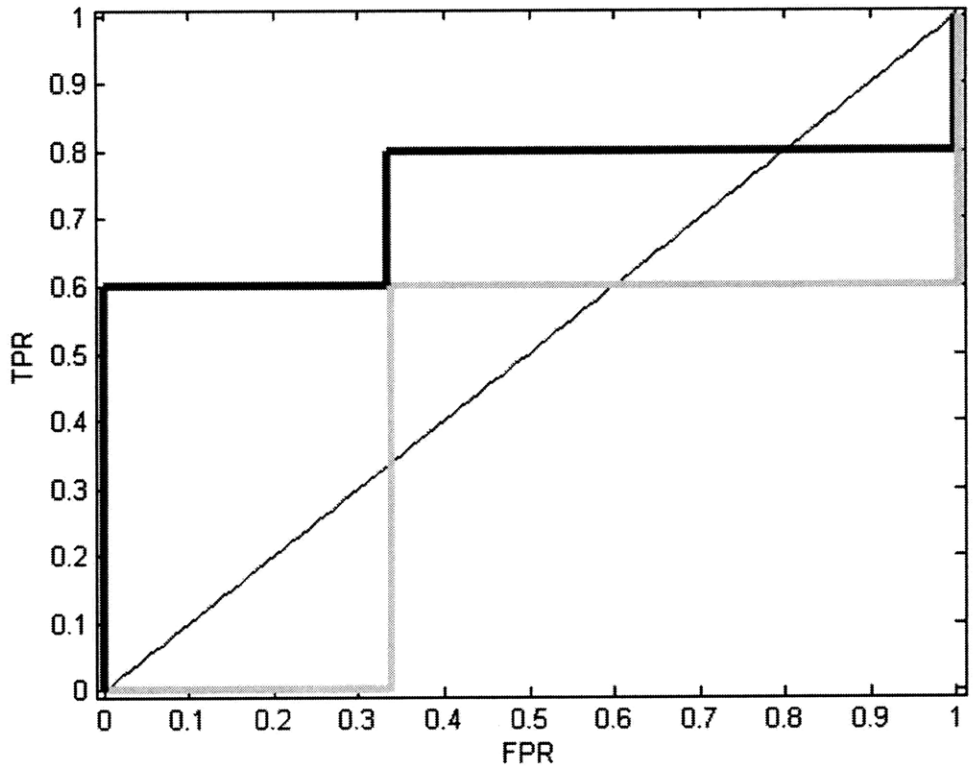
Supplementary Fig. 3B: Rotated view of Fig. 3A



Supplementary Fig. 3. RacGAP50C and RhoGAP93B single- and double-knockouts. **(A)** The plot in PC-based coordinates shows single-cell point sets for RacF28L (green), RacGAP50C single-knockout (magenta), RhoGAP93B single-knockout (cyan), and RacGAP50C/RhoGAP93B double-knockout (yellow). The classification model maps the RacGAP50C single-knockout to RacF28L with low confidence and maps the RhoGAP93B single-knockout to Rho1 with low confidence, but it maps the RacGAP50C/RhoGAP93B double-knockout to Rac1 with high confidence (correctly). **(B)** Rotated view, illustrating the extreme location of the point set for the RhoGAP93B single-knockout relative to the RacF28L point set.

Supplementary Figure 4

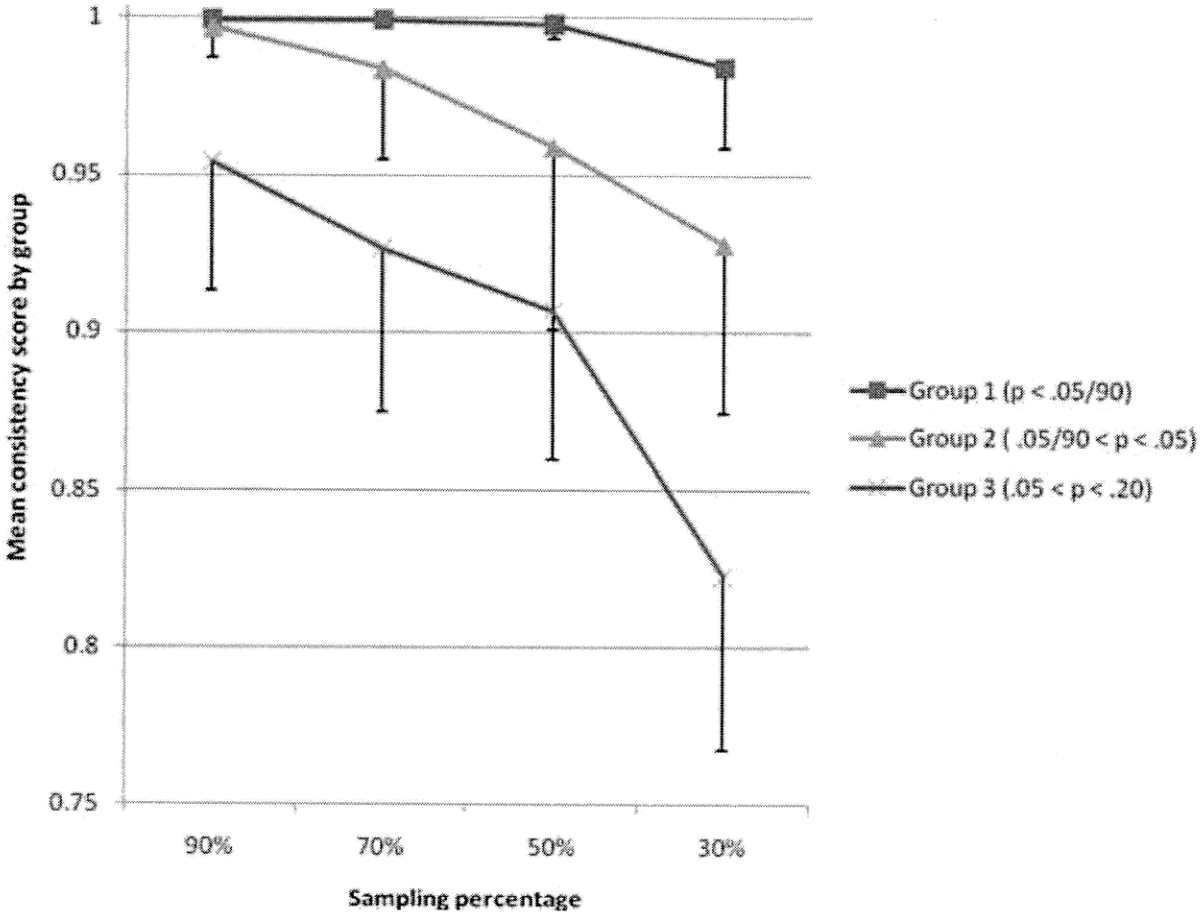
ROC curve for neural network based-alternative classification model



Supplementary Fig. 4. ROC curve for neural network-based alternative classification model. An alternative classification model was constructed using neural network classifiers for RhoF30L and RacF28L to perform dimensionality reduction, in contrast to dimensionality reduction based on principal components, as in the main classification model proposed in this paper). The figure shows the ROC curve for the main classification model (blue) and for the alternative model (green). The neural network-based model cannot achieve greater than 60% sensitivity, and can only do so at 67% specificity. The PC-based model outperforms it and all other alternatives considered (see main text for discussion of alternative methods).

Supplementary Figure 5

Robustness of classification to exclusion of data using jackknifing

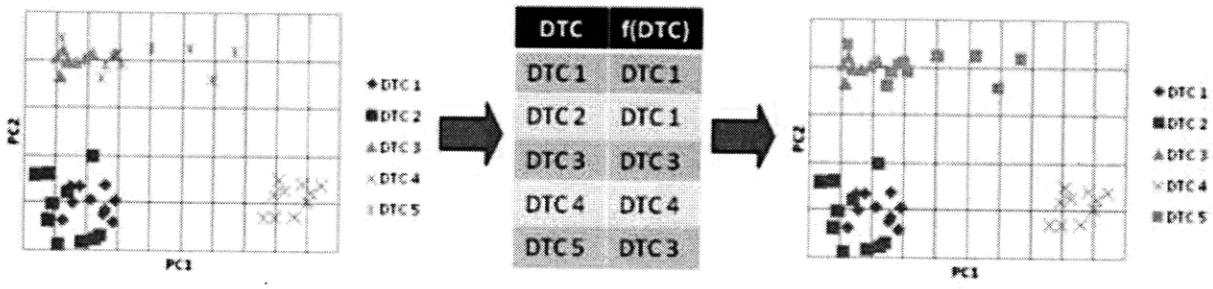


Supplementary Fig. 5. Robustness of classification to exclusion of data using jackknifing. For each single- and double-knockout, 100 random samples consisting of X% (X = 30, 50, 70, 90) of the cells from that TC were selected and classified to the set of overexpression TCs. A consistency score was assigned based on the fraction of random samples correctly classified. Single- and double-knockouts were binned into groups depending on p-score of the true classification. Mean and standard deviations of consistency scores are shown here for the three groups defined by largest p-scores (see figure legend). Most importantly, classifications of the

TCs that were classified the optimal threshold of $p = .0232$ are extremely robust to data exclusion (top line in graph). See **Supplementary Table 12** for full jackknife results.

Supplementary Figure 6

Classification-based clustering algorithm for downstream target TCs



Supplementary Fig. 6. Classification-based clustering algorithm for downstream target TCs. The first iteration ($k = 0, D_0 = D$) of the algorithm is shown for hypothetical data consisting of a set, D, of five downstream TCs (DTCs). On the left, the single cells for all DTCs are shown in reduced feature space. In the middle (step (i) of the algorithm), the classification model is used to map each of the DTCs to the set D_0. In this example, each of DTC1, DTC3, and DTC4 is mapped to itself, but the other two DTCs are not. On the right (step (ii) of the algorithm), DTCs classified to the same target are consolidated into growing clusters. For the next iteration of the algorithm ($k = 1$), each of the five original DTCs is classified, but this time the target set consists of only three targets, as $D_1 = \{DTC1 \cup DTC2, DTC3 \cup DTC5, DTC4\}$.

Supplementary Table 1

List of raw geometric features for *Drosophila* screens

Geometric Feature (*Drosophila*)
Area
Solidity
Eccentricity
MajorAxisLength
MinorAxisLength
EquivDiameter
MeanIntensity
StdIntensity
90thPercentileIntensity
GFPBrightSpotMajorSegments
GFPBrightSpotTotalArea
GFPBrightSpotMajorSegmentAreaMean
GFPBrightSpotMajorSegmentAreaCV
GFPBrightSpotMajorSegmentMaxMinSeparation
GFPBrightSpotGFPCentroidRelOffset
GFPCentroidGFPCenterOfMassRelOffset
GFPBrightSpotGFPCenterOfMassRelOffset
GFPCenterOfMassGFPMomentOfInertia
GFPCentroidGFPMomentOfInertia
GFPBrightSpotGFPMomentOfInertia
GFPMultivariateKurtosis
GFPHalfMassRelDistanceFromBoundary
GFPHalfMassRelDistanceFromGFPCenterOfMass
GFPHalfMassRelDistanceFromGFPCentroid
GFPHalfMassRelDistanceFromGFPBrightSpotCentroid
RuffleArea
RufflePixSum
RuffleVolume
DrainageArea
DrainagePixSum
GFPEdgeNumber
GFPEdgeTotalPixels
GFPEdgePixelDensity
GFPEdgeMeanLength
GFPEdgeMeanRelativeLength

GFPIntensityLocationMutualInformation_5_15_15
GFPIntensityLocationMutualInformation_8_15_24
GFPIntensityLocationMutualInformation_5_20_15
GFPIntensityLocationMutualInformation_8_20_24
GFPGauss2DFitMeanResidual
GFPGauss2DFitCorrelation
GFPGauss2DFitRelativeSigmaRow
GFPGauss2DFitRelativeSigmaCol
GFPGauss2DFitRelativeOffsetMeanFromSegCentroid
GFPGauss2DFitRelativeOffsetMeanFromBrightSpotCentroid
LoSmoothEccentricity
LoSmoothMajorAxisLength
LoSmoothMinorAxisLength
LoSmoothEllipticity
LoSmoothGFPCentroidClosestFocusRelOffset
LoSmoothGFPCenterOfMassClosestFocusRelOffset
LoSmoothGFPBrightSpotClosestFocusRelOffset
LoSmoothBndNormIntegratedAbsAngle
LoSmoothBndUndulationCount
LoSmoothBndUndulationTotalRelativeArea
LoSmoothBndProcessesGE0.5
LoSmoothBndProcessesGE1
LoSmoothBndCurvatureSharpestProcess
LoSmoothAreaSharpestProcess
LoSmoothRelativeAreaSharpestProcess
LoSmoothBndCurvature2ndSharpestProcess
LoSmoothArea2ndSharpestProcess
LoSmoothRelativeArea2ndSharpestProcess
LoSmoothBndAngleSharpestProcessesGFPCentroid
LoSmoothBndAngleSharpestProcessesGFPCenterOfMass
LoSmoothBndAngleSharpestProcessesGFPBrightSpotCentroid
LoSmoothHeightTallestProcess
LoSmoothRelativeHeightTallestProcess
LoSmoothAreaTallestProcess
LoSmoothRelativeAreaTallestProcess
LoSmoothBaseTallestProcess
LoSmoothRelativeBaseTallestProcess
LoSmoothHeight2ndTallestProcess
LoSmoothRelativeHeight2ndTallestProcess
LoSmoothArea2ndTallestProcess
LoSmoothRelativeArea2ndTallestProcess

LoSmoothBase2ndTallestProcess
LoSmoothRelativeBase2ndTallestProcess
LoSmoothBndAngleTallestProcessesGFPCentroid
LoSmoothBndAngleTallestProcessesGFPCenterOfMass
LoSmoothBndAngleTallestProcessesGFPBrightSpotCentroid
LoSmoothBndLargestAreaForProcessGE0.5
LoSmoothBndLargestRelativeAreaForProcessGE0.5
LoSmoothBndSecondLargestAreaForProcessGE0.5
LoSmoothBndSecondLargestRelativeAreaForProcessGE0.5
LoSmoothBndAngleLargestProcessesGE0.5GFPCentroid
LoSmoothBndAngleLargestProcessesGE0.5GFPCenterOfMass
LoSmoothBndAngleLargestProcessesGE0.5GFPBrightSpotCentroid
LoSmoothBndLargestAreaForProcessGE1
LoSmoothBndLargestRelativeAreaForProcessGE1
LoSmoothBndSecondLargestAreaForProcessGE1
LoSmoothBndSecondLargestRelativeAreaForProcessGE1
LoSmoothBndAngleLargestProcessesGE1GFPCentroid
LoSmoothBndAngleLargestProcessesGE1GFPCenterOfMass
LoSmoothBndAngleLargestProcessesGE1GFPBrightSpotCentroid
HiSmoothEccentricity
HiSmoothMajorAxisLength
HiSmoothMinorAxisLength
HiSmoothEllipticity
HiSmoothGFPCentroidClosestFocusRelOffset
HiSmoothGFPCenterOfMassClosestFocusRelOffset
HiSmoothGFPBrightSpotClosestFocusRelOffset
HiSmoothBndNormIntegratedAbsAngle
HiSmoothBndUndulationCount
HiSmoothBndUndulationTotalRelativeArea
HiSmoothBndProcessesGE0.5
HiSmoothBndProcessesGE1
HiSmoothBndCurvatureSharpestProcess
HiSmoothAreaSharpestProcess
HiSmoothRelativeAreaSharpestProcess
HiSmoothBndCurvature2ndSharpestProcess
HiSmoothArea2ndSharpestProcess
HiSmoothRelativeArea2ndSharpestProcess
HiSmoothBndAngleSharpestProcessesGFPCentroid
HiSmoothBndAngleSharpestProcessesGFPCenterOfMass
HiSmoothBndAngleSharpestProcessesGFPBrightSpotCentroid
HiSmoothHeightTallestProcess

HiSmoothRelativeHeightTallestProcess
HiSmoothAreaTallestProcess
HiSmoothRelativeAreaTallestProcess
HiSmoothBaseTallestProcess
HiSmoothRelativeBaseTallestProcess
HiSmoothHeight2ndTallestProcess
HiSmoothRelativeHeight2ndTallestProcess
HiSmoothArea2ndTallestProcess
HiSmoothRelativeArea2ndTallestProcess
HiSmoothBase2ndTallestProcess
HiSmoothRelativeBase2ndTallestProcess
HiSmoothBndAngleTallestProcessesGFPCentroid
HiSmoothBndAngleTallestProcessesGFPCenterOfMass
HiSmoothBndAngleTallestProcessesGFPBrightSpotCentroid
HiSmoothBndLargestAreaForProcessGE0.5
HiSmoothBndLargestRelativeAreaForProcessGE0.5
HiSmoothBndSecondLargestAreaForProcessGE0.5
HiSmoothBndSecondLargestRelativeAreaForProcessGE0.5
HiSmoothBndAngleLargestProcessesGE0.5GFPCentroid
HiSmoothBndAngleLargestProcessesGE0.5GFPCenterOfMass
HiSmoothBndAngleLargestProcessesGE0.5GFPBrightSpotCentroid
HiSmoothBndLargestAreaForProcessGE1
HiSmoothBndLargestRelativeAreaForProcessGE1
HiSmoothBndSecondLargestAreaForProcessGE1
HiSmoothBndSecondLargestRelativeAreaForProcessGE1
HiSmoothBndAngleLargestProcessesGE1GFPCentroid
HiSmoothBndAngleLargestProcessesGE1GFPCenterOfMass
HiSmoothBndAngleLargestProcessesGE1GFPBrightSpotCentroid

Supplementary Table 1. List of raw geometric features (145) for *Drosophila* screens. See [10] for additional information about raw data acquisition.

Supplementary Table 2
List of GAPs included in genetic screen

GAPs Included in Double-Knockout Screen
CdGAPr
RhoGAP100F
RhoGAP16F
RhoGAPp190
RhoGAP19D
RhoGAP1A
RacGAP50C
RhoGAP54D
RhoGAP5A
RhoGAP71E
RacGAP84C
RhoGAP92B
RhoGAP93B

Supplementary Table 2. List of GAPs included in genetic screen. All single-knockouts and all possible combinations of double-knockouts except for RhoGAP19D/RhoGAP54D were included in the screen, for a total of 90 distinct TCs. In all, 6480 single cells were imaged across these TCs.

Supplementary Table 3

Biologically validated RhoGAP/GTPase interactions and non-interactions

Supplementary Table 3A: Biologically-validated interactions

GAP	GTPase	Reference
RacGAP50C	Rac1	Sotillos and Campuzano, 2000
RacGAP84C	Rac1	Raymond et al., 2001
RhoGAP93B	Rac1	Lundstrom et al., 2004
RhoGAPp190	Rho1	Billuart et al., 2001
RacGAP50C	Cdc42	Sotillos and Campuzano, 2000

Supplementary Table 3A. Biologically validated RhoGAP/GTPase interactions among the 13 GAPs and 3 GTPases in our morphological datasets, taken from Flybase and BioGRID. Relevant citations are provided in the third column.

Supplementary Table 3B: Biologically-validated non-interactions

GAP	GTPase	Reference
RhoGAP5A	Rho1	{cite}
RacGAP84C	Rho1	{cite}
CdGAPr	Rho1	{cite}

Supplementary Table 3B. Biologically validated RhoGAP/GTPase non-interactions among the 13 GAPs and 3 GTPases in our morphological datasets.

Supplementary Table 4

Classification of single-knockout GAP TCs into GTPase overexpression TCs

GAP	GTPase	P-Score
RhoGAP92B	Rho1	0.000
RacGAP84C	Rho1	0.011
RhoGAPp190	Rho1	0.021
RhoGAP19D	Rho1	0.046
CdGAPr	Rac1	0.167
RhoGAP54D	Rac1	0.205
RacGAP50C	Rac1	0.231
RhoGAP93B	Rho1	0.437
RhoGAP5A	Cdc42	0.542
RhoGAP16F	Rho1	0.657
RhoGAP1A	Rho1	0.701
RhoGAP100F	Rac1	0.891
RhoGAP71E	Rho1	0.980

Supplementary Table 4. Classification of single-knockout GAP TCs into GTPase overexpression TCs. The confidence scores shown here were computed using bootstrapping by drawing samples just from the 13 single-knockout GAP 13 TCs. Following Bonferroni correction, only the mapping of RhoGAP92B to Rho1 is significant at $p = .05$ (heavy shading). By considering the ROC curve (**Fig. 2D**), this model has optimal predictive power at a threshold of $p = .231$ (light shading), at which it correctly predicts 2/5 biologically-validated interactions and 2/3 non-interactions. At this threshold, the model makes a total of 7 predictions. The probability of correctly predicting at least 2 out of 5 biologically-validated interactions when making 7 predictions (out of 39 possibly) is $p = .21$.

Supplementary Table 5

Classification of single- and double-knockout GAP TCs into GTPase overexpression TCs

Supplementary Table 5A: Classifications

	Cd.r	100F	16F	p190	19D	1A	50C	54D	5A	71E	84C	92B	93B
Cd.r	Rac1	Rho1	Rho1	Rho1	Rho1	Rac1	Rac1	Cdc42	Rac1	Rho1	Rac1	Rho1	Rho1
100F		Rac1	Cdc42	Rho1	Rho1	Rho1	Rac1	Rho1	Rho1	**Rac1**	Rho1	Rho1	Rho1
16F			Rho1	Rho1	Rho1	Rac1	**Rac1**	Cdc42	Cdc42	Rac1	Rho1	Rho1	Rho1
p190				**Rho1**	Rho1	Rho1	Rac1	Rho1	Cdc42	Rho1	Rho1	Rho1	Rho1
19D					**Rho1**	Rho1	Rac1	n/a	Rac1	Rac1	Rho1	Rho1	Rho1
1A						Rho1	**Rac1**	Cdc42	Rho1	Rac1	Rho1	Rho1	Rho1
50C							Rac1	**Rac1**	Rac1	**Rac1**	**Rac1**	Rho1	**Rac1**
54D								Rac1	Rac1	Rac1	Rho1	Rho1	Rac1
5A									Cdc42	Rac1	Rho1	Rho1	Rho1
71E										Rho1	Rho1	Rho1	Rho1
84C											**Rho1**	Rho1	Rac1
92B												**Rho1**	Rho1
93B													Rho1

Supplementary Table 5B: P-scores for classifications

	Cd.r	100F	16F	p190	19D	1A	50C	54D	5A	71E	84C	92B	93B
Cd.r	0.055	0.789	0.947	0.078	0.851	0.500	0.601	0.941	0.197	0.584	0.988	0.010	0.725
100F		0.749	0.103	0.869	0.112	0.405	0.026	0.822	0.360	**0.006**	0.349	0.003	0.312
16F			0.368	0.278	0.157	0.984	**0.000**	0.860	0.621	0.424	0.235	0.000	0.998
p190				**0.001**	0.854	0.886	0.047	0.410	0.743	0.316	0.174	0.000	0.681
19D					**0.012**	0.092	0.175	n/a	0.666	0.607	0.122	0.000	0.107
1A						0.505	**0.000**	0.541	0.635	0.109	0.274	0.003	0.846
50C							0.086	**0.000**	0.135	**0.000**	**0.000**	0.000	**0.023**
54D								0.091	0.087	0.468	0.239	0.000	0.643
5A									0.309	0.143	0.517	0.000	0.747
71E										0.965	0.186	0.146	0.313
84C											**0.003**	0.000	0.958
92B												**0.000**	0.000
93B													0.213

Supplementary Table 5. **(A)** Classification of single- and double-knockout GAP TCs into GTPase overexpression TCs. **(B)** P-scores associated with classifications, as determined by bootstrapping with 1000 iterations. Classifications significant at $p = .0232$ are lightly shaded. Double-knockouts with RhoGAP92B as one of the knocked-down genes (the heavily shaded cells in the upper right of the matrix) were excluded, due to the fact that RhoGAP92B single-knockout was mapped to Rho1 overexpression at Bonferroni-corrected $p = .05$ (see **Supplementary Table 4**). The threshold of $p = .0232$ was chosen based on ROC analysis to yield maximum sensitivity while simultaneously minimizing the false positive rate (see **Fig. 2D**).

Supplementary Table 6

Clustering of single-knockout GAP TCs

Single-knockout GAP TC	Cluster Index
CdGAPr	1
RhoGAP92B	2
RhoGAPp190	3
RhoGAP19D	3
RacGAP84C	3
RhoGAP100F	4
RhoGAP16F	4
RhoGAP5A	4
RhoGAP93B	4
RhoGAP1A	5
RacGAP50C	5
RhoGAP54D	5
RhoGAP71E	5

Supplementary Table 6. Clustering of single-knockout GAP TCs. A clustering algorithm was designed was used to ensure that, following clustering, each single-knockout GAP TC would be mapped to the cluster containing it under the classification model. See text for details.

Supplementary Table 7
Classification of double-knockout GAP TCs into single-knockout GAP TCs

Supplementary Table 7A: Classifications

	Cd.r	100F	16F	p190	19D	1A	50C	54D	5A	71E	84C	92B	93B
Cd.r	1	5	5	4	5	5	5	4	5	5	5	5	5
100F		4	**4**	4	3	5	4	5	4	5	3	2	5
16F			4	4	5	5	5	4	1	1	2	**2**	4
p190				3	4	4	5	4	4	5	4	**2**	4
19D					3	5	5	n/a	5	5	4	2	1
1A						5	5	4	4	5	5	2	5
50C							5	**5**	5	**5**	**5**	3	5
54D								**5**	4	5	4	2	5
5A									**4**	5	4	**2**	4
71E										5	5	5	**5**
84C											3	2	5
92B												**2**	3
93B													4

Supplementary Table 7B: P-scores for classifications

	Cd.r	100F	16F	p190	19D	1A	50C	54D	5A	71E	84C	92B	93B
Cd.r	0.192	0.676	0.022	0.888	0.003	0.045	0.195	0.064	0.008	0.548	0.071	0.260	0.073
100F		0.277	**0.000**	0.795	0.502	0.778	0.059	0.820	0.798	0.052	0.756	0.911	0.485
16F			0.631	0.081	0.418	0.259	0.005	0.092	0.069	0.159	0.789	**0.000**	0.506
p190				0.300	0.491	0.429	0.446	0.792	0.097	0.031	0.640	**0.000**	0.092
19D					0.146	0.700	0.002	n/a	0.129	0.299	0.160	0.311	0.905
1A						0.597	0.014	0.005	0.538	0.001	0.824	0.283	0.334
50C							0.113	**0.000**	0.010	**0.000**	**0.000**	0.592	0.213
54D								**0.000**	0.012	0.464	0.489	0.888	0.002
5A									**0.000**	0.364	0.219	**0.000**	0.141
71E										0.013	0.673	0.791	**0.000**
84C											0.895	0.017	0.390
92B												**0.000**	0.265
93B													0.261

Supplementary Table 7. (A) Classification of double-knockout GAP TCs into single-knockout GAP TCs. The code (1-5) corresponds to the cluster index given in **Supplementary Table 6**. **(B)** P-scores associated with classifications, as determined by bootstrapping with 1000 iterations. Classifications significant after Bonferroni correction at $p = .05$ are shaded. For discussion, see main text and **Fig. 3**.

Supplementary Table 8

Robustness of classification to method of dimensionality reduction

	1 PC	2 PCs	3 PCs	4 PCs	5 PCs	6 PCs	7 PCs	8 PCs	9 PCs	10 PCs
Prob.	2E-9	6E-13	<1E-14	7E-14	2E-11	3E-9	9E-3	3E-3	2E-4	5E-4

Supplementary Table 8. Robustness of classification to method of dimensionality reduction. The entire set of GAP single- and double-knockouts TCs was reclassified to the set of GTPase overexpression TCs by using varying numbers of principal components for the dimensionality reduction. Bootstrapping was performed with 100 iterations to determine p-scores for classifications. The hypergeometric distribution was used to calculate the probability of achieving equal or greater overlap of sets of TCs scoring at Bonferroni-corrected $p = .05$ for each classification under alternate dimensionality reduction, as compared to the overlap observed with the actual classification. Observed overlap is highly significant using six PCs or fewer, indicating that results are robust to the method of dimensionality reduction in this range. By increasing the number of principal components, additional noise is introduced into the data representing the single cell populations for each TC, thus complicating classification to the GTPase overexpression TCs (the GTPase overexpression TCs are clearly distinguishable from one another using just three principal components; see **Supplementary Table 10**).

Supplementary Table 9

Alternative bootstrapping for mapping single-knockout GAP TCs to GTPase overexpression TCs

GAP	GTPase	P-Score (Bootstrapping from single-knockouts only)	P-Score (Bootstrapping from single- and double-knockouts)
RhoGAP92B	Rho1	0.000	0.000
RacGAP84C	Rho1	0.011	0.003
RhoGAPp190	Rho1	0.021	0.001
RhoGAP19D	Rho1	0.046	0.012
CdGAPr	Rac1	0.167	0.055
RhoGAP54D	Rac1	0.205	0.091
RacGAP50C	Rac1	0.231	0.086
RhoGAP93B	Rho1	0.437	0.213
RhoGAP5A	Cdc42	0.542	0.309
RhoGAP16F	Rho1	0.657	0.368
RhoGAP1A	Rho1	0.701	0.505
RhoGAP100F	Rac1	0.891	0.749
RhoGAP71E	Rho1	0.980	0.965

Supplementary Table 9. Alternative bootstrapping for mapping single-knockout GAP TCs to GTPase overexpression TCs. P-scores were calculated for classifications using two alternate methods: bootstrapping by drawing samples from single-knockout TCs only (third column) and by drawing samples from both single- and double-knockout TCs (fourth column). Sampling from both single- and double-knockouts results in uniformly smaller p-scores, because including double-knockouts increases the range of single-cell morphological phenotypes available. The ordering of the 13 GAPs by p-score was maintained by shifting between the two bootstrapping schemes, save for a swap in the order of two adjacent pairs of interactions; the ROC curve is unchanged overall.

Supplementary Table 10

Classification of the set of GTPase overexpression TCs to itself

"Upstream" TC	Classification ("Downstream" TC)	P-Score
RhoF30L	RhoF30L	0.000
RacF28L	RacF28L	0.000
Cdc42Y32A	Cdc42Y32A	0.000

Supplementary Table 10. Classification of the set of GTPase overexpression TCs to itself. Each of the three GTPase overexpression TCs is correctly mapped. P-scores were calculated using bootstrapping with 1000 iterations, with samples drawn from the union (of single cells) of the three GTPase overexpression TCs. The fact that all p-scores were 0.000 means that, for all three TCs, no bootstrapped samples of equal cell number were classified with greater confidence (mode frequency of the classification vector) than the actual set of cells. This provides significant confidence in the ability of the classification model to discriminate between the GTPase overexpression TCs. (In contrast, a preliminary clustering step was necessary when mapping (double-knockout GAP TCs) to the set of single-knockout GAP TCs).

Supplementary Table 11

Robustness of classification to exclusion of data using jackknifing

Supplementary Table 11A: Groupings for single- and double-knockouts based on P-score

Group Definition	Group Size
Group 1 ($p < .05/90$)	14
Group 2 ($.05/90 < p < .05$)	10
Group 3 ($.05 < p < .20$)	19
Group 4 ($.20 < p < .50$)	16
Group 5 ($.50 < p < .80$)	16
Group 6 ($.80 < p$)	15
Total	**90**

Supplementary Table 11B: Mean consistency scores by grouping

	90%	70%	50%	30%
Group 1 ($p < .05/90$)	0.999 (0.003)	0.999 (0.003)	0.998 (0.004)	0.984 (0.026)
Group 2 ($.05/90 < p < .05$)	0.997 (0.010)	0.984 (0.029)	0.959 (0.058)	0.928 (0.054)
Group 3 ($.05 < p < .20$)	0.954 (0.041)	0.927 (0.052)	0.907 (0.047)	0.823 (0.055)
Group 4 ($.20 < p < .50$)	0.856 (0.078)	0.813 (0.080)	0.784 (0.073)	0.735 (0.065)
Group 5 ($.50 < p < .80$)	0.724 (0.112)	0.707 (0.112)	0.679 (0.102)	0.598 (0.098)
Group 6 ($.80 < p$)	0.566 (0.116)	0.532 (0.096)	0.511 (0.077)	0.499 (0.053)

Supplementary Table 11. Robustness of classification to exclusion of data using jackknifing. For each single- and double-knockout, 100 random samples consisting of X% (X = 30, 50, 70, 90) of the cells from that TC were selected and classified to the set of overexpression TCs. A consistency score was assigned based on the fraction of random samples correctly classified. **(A)** Single- and double-knockouts were binned into groups depending on p-score of the true classification. **(B)** Mean consistency scores are shown here for all groups (standard deviations are shown in parentheses). Most importantly, classifications of the TCs that were classified at

the optimal threshold of $p = .0232$ are extremely robust to data exclusion (top two lines in table). See also **Supplementary Fig. 5.**

Supplementary Table 12

Sensitivity/specificity for mapping single- and double-knockout GAP TCs to GTPase overexpression TCs

	Biologically validated positive	**Biologically validated negative**
Classified positive by model	4	1
Classified negative by model	1	2

Supplementary Table 12. Sensitivity/specificity for mapping single- and double-knockout GAP TCs to GTPase overexpression TC at optimal threshold ($p = .0232$; see **Fig. 2D**). The model identified 2 true positive interactions (RacGAP50C/Rac1, RacGAP84C/Rac1), 3 false positives, 0 false negatives, and 3 true negatives, yielding a sensitivity of 4/5 (80%) and specificity of 2/3 (67%). At this threshold, the model makes a total of 12 predictions (see **Fig. 2**). Furthermore, by hypergeometric statistics, the probability of correcting predicting at least 4 out of 5 biologically-validated interactions when making 12 predictions (out of 39 total possible) is $p = .0246$. This supports the significance of our model, especially in comparison to alternatives which make many more positive predictions, thus decreasing the overall predictive power of the model (even if sensitivity and specificity scores are comparable).

Chapter 4:
Data integration for High-throughput Morphological and Transcriptional Genetic Screens

Abstract

A recurrent theme in computational biology has been the development of methods to combine different data sources for increased predictive power. With the emergence of high-throughput morphological screens, a key challenge is to integrate this data source with high-throughput transcriptional data from microarrays. Here, we apply techniques from microarray data analysis to determine differential expression and gene set enrichment between group pairs defined by rigorous, quantitative morphology-based class distinctions. By comparing expression data between control treatment conditions and treatment conditions displaying a particular morphological phenotype of interest (e.g. high single-cell morphological variability or inability to form lamellipodia), we identify genes and pathways correlated with this class distinction. We apply these techniques to morphological data from a genetic screen using *Drosophila* BG-2 cells, microarray data from a screen in *Drosophila* S2R+ cells, and morphological class distinctions defined in several previous studies. We identify meaningful differential expression or pathway/functional category enrichment for several morphological class distinctions, thus highlighting putative mechanisms for morphological change and generating new genes of

interest for future study. Based on the success of this study, we expect that these techniques will prove useful in further data integration studies.

Introduction

A recurrent theme in computational biology has been the development of methods to combine different data sources for increased predictive power. The use of disparate data sources carries with it multiple advantages, chief among them enhanced ability to detect phenotypes (increased sensitivity) and improved ability to double-check, in essence, the predictions of one data source against another (increased specificity). Examples highlighting the power of the integration of multiple data sources include PPI alignment [1-3], enhancer prediction [4-6], transcriptional network inference [7-10] and signaling pathway inference [11]. With these powerful examples of data integration as motivation, we propose a framework for combining two important data sources –high-throughput transcriptional data and high-throughput morphological data from genetic screens – in order to gain increased understanding of the genetics of morphological phenotypes.

The acquisition of high-throughput morphological data has matured in recent years [12, 13]. Techniques for using this data to identify phenotypes have been developed in various contexts, to quantify shape [14], DNA morphology [15], and subcellular-localization of organelles or proteins [16, 17], on a single-cell level. Initial analysis was commonly performed by averaging single-cell results to derive mean scores or by clustering such results [14, 18-20]. Recently, researchers have quantified morphological variability on the single-cell level in response to various stimuli, e.g. genetic or chemical perturbations [21-24]. In this work, we focus on

morphological phenotypes corresponding to clustering of mean morphologyand quantification of population-level variability. We seek to study the mechanisms behind these phenomena by integrating analysis of morphological data with expression data. The use of morphology to define class distinctions for further study by transcriptional data has a long history, particularly in the study of cancer [25], but has not been applied, to our knowledge, to high-throughput morphological data from a large-scale genetic screen.

The literature on the analysis of transcriptional data is well-developed. Most relevant for our work are methods to detect differential expression between two unpaired groups of treatment conditions [25-27]. A complication in microarray data analysis has been the proliferation of alternative methods for data normalization, values for cutoff parameters, and, more basically, methods for determining significant differential expression. In this study, we use two alternatives for determining differential expression, one of which, t tests, is essentially the "industry standard." However, the results for differential expression of individual genes are limited: we find few genes to be differentially expressed across the morphological class distinctions under consideration. Accordingly, we make use of gene set enrichment analysis techniques for our expression data [28-29]. These techniques are able to detect overall enrichment for a gene set, even if the individual differential expression of the component genes is not statistically significant.

Our main aim in this paper is to establish a framework for integrating high-throughput data from morphological and transcriptional genetic screens in order to study genes and pathways involved in different morphological phenotypes. An outline of the approach that we adopt is as follows. We first acquire high-throughput single-cell morphological data from a genetic screen, meaning that we create multiple experimental treatment conditions (TCs), each of which is defined by

knockout or overexpression of a particular gene or genes, and use image processing techniques to acquire the single-cell data [14]. In parallel, we acquire high-throughput transcriptional data by running microarrays on the same TCs (**Supplementary Table 1**) in the morphological genetic screen [30]. In our case, we work in *Drosophila*, with the caveat that the morphological data and transcriptional data were obtained from two different cell lines (BG-2 for morphological; S2R+ for transcriptional). Using this data, we define a number of class distinctions on the basis of differences in morphology. An example of a class distinction that we considered is high population morphological variability versus low population morphological variability. Each class distinction permits us to define two groups of TCs, each group corresponding to one side or the other of the distinction. We then turn to the microarray data, on the basis of which we study differential expression and gene set enrichment between these two groups of TCs.

Whether changes in expression of the genes identified in this manner are causal of or are caused by the morphological phenotype must be investigated by further experimentation. However, we do use previous results in the literature for the genes and pathways that we identify in order to validate our findings and provide support for our framework of study. Overall, our results provide insights into the mechanisms for morphological change. Based on the progress reported here, we strongly recommend that experiments be carried out to obtain expression data in the BG-2 cell line, which will permit much more precise investigation of mechanisms involved in the morphological class distinctions under consideration.

Results

We systematically studied differences in expression between TCs showing different morphologies using the following fundamental framework (**Fig. 1**). After processing morphological and transcriptional data from the genetic screens (**Materials and Methods**), two distinct steps were required. First, we defined a number of class distinctions in order to separate treatment conditions into groups of classes on the basis of morphology. We generated two different types of class distinctions corresponding to phenocluster analysis and variability analysis (**Supplementary Table 2**). Second, we applied computational tools to determine differential expression and gene set enrichment between all pairs of TC groups (**Materials and Methods**).

Differential expression

As a first step to study differential expression, we performed t-tests (**Materials and Methods**) on processed microarray data according to each of the class distinctions that we defined. However, the number of individual genes identified as being significantly differentially expressed was minimal in all cases (**Table 1**). The likely reasons for this are twofold. First, the quality of the microarray data set is relatively poor in thatthe range of intensity values for expression is smaller than that obtained for typical microarray studies. Second, the fact that we use different cell lines means that the signal strength is dampened. For example, a treatment condition displaying morphology of a particular phenocluster in the BG-2 cell line may not cluster with the same TCs in the S2R+ cell line.

Consequently, we chose to carry out Significance Analysis of Microarrays (SAM, see **Materials and Methods**), which is a less stringent method for determining differential expression [31]. By

applying a consistent FDR cutoff, we were able to study differential expression systematically for each of the morphological class distinctions (**Table 1**). Note that two of the class distinctions resulted in far more significantly differentially expressed genes than the other comparisons-namely, Low versus High Variability and Control versus the Adhesion Disassembly/Cortical Tension Phenocluster.

Since the number of individual genes identified by SAM was in some cases large, for biological validation we focused on the top 5 up/down-regulated genes (as determined by fold change, or FC) among the genes with FDR less than 5% (**Table 2** and **Fig. 2**).

Differential expression: Morphological variability

For the morphological class distinctions for Low versus High Variability, Control versus High Variability, and Control versus Low Variability, the first two comparisons yielded significant results at FDR of 5%, while the third comparison did not. The lack of results for the Control versus Low Variability comparison may suggest that expression is highly similar for the control and low variability TC groups. However, this simple conclusion is challenged by the fact that many more genes were differentially expressed for the Low versus High Variability comparison than for the Control versus High Variability comparison.

A number of biologically meaningful genes, in the context of morphological variability, were highlighted by SAM analysis (**Table 2**). CG30440, which encodes a RhoGEF, had the largest FC among the genes up-regulated in the Control versus High Variability comparison. This result is validated by the fact that knockout of CG30440 resulted in a single-cell population with low population variability, as measured by variability p-scores [21]. Note that the CG30440 TC had

a variability p-score of marginal significance ($p < .05$ before, but not after correction for multiple hypotheses), which highlights the ability of data integration to identify novel genes involved in morphological phenotypes. Also of note, the gene Ef2b, which encodes a protein involved in translation elongation and which has been implicated in regulation of the mitotic spindle, was up-regulated in the Low versus High Variability comparison [32, 33]. Several kinases and phosphatases (gp150, Tao-1, Protein tyrosine phosphatase 4E) were found to be up-regulated in the High Variability TC group, a finding which provides further motivation to pursue gene set analysis. Note additionally that several genes of unknown function were found to be differentially expressed (bcn92 and fok were up-regulated in the High Variability group; wibg was up-regulated in the Low Variability group), which illustrates the ability of data integration to generate new targets for further biological experimentation.

Differential expression: Phenoclusters

For the morphological class distinctions for control versus each of the six phenoclusters, respectively, our differential expression findings were consistent with known properties of the genes in the expression dataset and the biological properties defining the phenoclusters.

For the Rac1 Phenocluster and the Rho1 Phenocluster, comparison with control identified Rac1 and Rho1, respectively, as the most down-regulated genes. This was an expected result, since the Rac1 and Rho1 Phenoclusters consist of TCs with similar morphology to the Rac1_RNAi and Rho1_RNAi TCs. In fact, the TC groups for each of these comparisons consist of several Rac1_RNAi and Rho1_RNAi microarray replicates, respectively (see **Supplementary Table 2**). The comparison of Control versus the Rho1 Phenocluster yielded no further results, whereas

Control versus the Rac1 Phenocluster identified the transcription factor tiptop as being down-regulated in the Rac1 cluster and the Ras GEF CG4853 as being up-regulated.

The comparison of Control versus the Lamellipodia Formation Phenocluster yielded more extensive results. The gene CG8636 was down-regulated in the Lamellipodia Formation Phenocluster TC group; this gene is involved in mitotic spindle elongation and translation [34]. In addition, two GTPases (Rab40, CG8641), a phosphatase (CG11597), and two genes of unknown function were found to be up-regulated.

By far the richest comparison among the control versus phenocluster morphological class distinctions, at least for the differential expression analysis, was for Control versus the Adhesion Disassembly/Cortical Tension Phenocluster (see **Table 1**). Remarkably, two of the same genes found to be up-regulated in the Lamellipodia Formation Phenocluster were among the top 5 (ranked by FC) of the 103 genes found to be differentially expressed in the Adhesion Disassembly/Cortical Tension Phenocluster. One of these genes was CG30220, a gene of unknown function, and the other was CG11597, a PP4-type phosphatase that has not been fully studied [35]. In addition, several genes involved in translation (Ef1gamma, eIF-5C, Trip1) were down-regulated in the Adhesion Disassembly/Cortical Tension Phenocluster (see gene set analysis, below). Hip1 (Huntingtin interacting protein 1) was identified as being up-regulated; this gene has recently been implicated in *Drosophila* neurogenesis [36]. In addition, the GTPase Sar1 was found to be down-regulated; Sar1 is involved in vesicle trafficking and has been shown to be essential in establishing cell polarity in neuronal cells [37].

No significant results were obtained for certain comparisons (namely for Protrusion/Adhesion Formation, and for the Rho1 Phenocluster, other than Rho1 itself, as discussed above). Several

genes were found to be differentially expressed in the comparison of GFP to other WT-appearing TCs (see **Table 2**), which points to noise in this approach (see **Discussion**). Overall, as with the variability-related class distinctions, multiple genes of unknown function were identified as being differentially expressed (see **Table 2**), again pointing to avenues for further research.

Gene set enrichment

In order to make better sense of the large number of significant genes identified by our differential expression analysis, we performed Gene Set Enrichment Analysis (GSEA, see **Materials and Methods**), using gene sets derived from KEGG [38] and GO [39] with 1000 permutations and an FDR threshold of 33 $\frac{1}{3}$%. Several of the class distinctions defined TC groups that are enriched for gene sets (**Table 3** and **Figs. 3** and **4**), yielding biologically meaningful results.

Gene set enrichment: Morphological variability

Among the three class distinctions related to variability, two of them yielded significant GSEA results. In particular, the Control versus High Variability comparison identified the ErbB pathway and the mTOR pathway as each being up-regulated in the High Variability group relative to control. And the Control versus Low Variability comparison identified the VEGF pathway and the cell cycle and gastrulation GO categories as being down-regulated in the Low Variability group. (Note that being up-regulated in the High Variability group relative to Control

and being down-regulated in the Low Variability group relative to Control are similar, but subtly different properties).

The ErbB and mTOR signaling pathways were both up-regulated in the high variability TC group relative to control (**Fig. 3**). This finding is consistent with the roles of these signaling pathways in regulation of cell locomotion, along with the fact that cells undergoing locomotion display higher population-level morphological variability [40, 41]. Activation of the ErbB signaling pathway has been extensively implicated in metastatic cancer [42]. Of note, previous work has studied the ability of the ErbB family for producing a diversity of signaling outputs due to multiple ligands and multiple receptors initiating distinct pathways in a combinatorial manner [43]. We extend this result by showing here that up-regulation of ErbB is correlated with high morphological variability. Likewise, the mTOR pathway is strongly associated with translational control, stress response, and locomotion [44, 45]. To determine whether up-regulation of the mTOR pathway is the cause of high population-level morphological variability, or whether it is a stress response requires further experimentation.

Gene sets for the VEGF signaling pathway and the GO category for gastrulation were found to be down-regulated in the low variability group relative to control. The VEGF pathway contains components related to regulation of the actin cytoskeleton, focal adhesion turnover, cell migration, and angiogenesis (among others) [46, 47]. The heightened activity of adhesion formation was earlier found to be correlated with high population variability [21]. Further, a recent paper has implicated both the VEGF and mTOR signaling pathways in regulation of cell size in *Drosophila* [48]; our result that these pathways are up-regulated in high variability TCs (or down-regulated in low variability TCs) provides an additional layer of complexity to those findings. For the case of gastrulation, previous studies have implicated Rho1 and RhoGEF2 in

the control of gastrulation in *Drosophila* embryos, and gastrulation involves coordinated cell movement and morphological changes [49-51]. Our finding that the gastrulation category is down-regulated in the low variability group likely reflects the overlap of certain key genes in regulation of gastrulation, on the one hand, and actin and motility regulation, on the other.

The finding that genes involved in cell cycle regulation are up-regulated in the Low Variability group is certainly interesting, but difficult to explain. How does up-regulation of these genes affect progression through the cell cycle? One conceivable possibility is that it speeds up cell cycle progression. But if this is the case, then there is no net effect on population-level morphological variability (i.e. the distribution of cell progress through the cycle would not change). On the other hand, if the pattern of up-regulation of this gene set results in,cell cycle arrest (even in a relatively small proportion of cases), then this could significantly decrease population-level morphological variability.

Gene set enrichment: Phenoclusters

Among the class distinctions for the six Phenoclusters, the richest GSEA results came from the Control versus Lamellipodia Formation Phenocluster comparison (**Fig. 4**). Four gene sets were found to be down-regulated at $q < .05$ (in general a cutoff of 33 $\frac{1}{3}$% was used to determine significance for GSEA). These gene sets were the Wnt pathway, VEGF pathway, cell cycle category, and gastrulation category. Note that three of these four sets were also down-regulated in the Low Variability group, indicating a point of similarity between the inability to form lamellipodia and low population-level morphological variability. The Wnt pathway is known to play a key role in neural crest migration and has been shown to direct formation of lamellipodia

[52]. At $q < .10$, GO categories for actin cytoskeleton and axonogenesis were identified as down-regulated for the Lamellipodia Formation Phenocluster; in the former case, this is consistent with the actin-based structure of lamellipodia. Multiple categories associated with translation were down-regulated, as were categories associated with development (dorsal closure, head involution, heart development, oogenesis), GTPase activity, GPCR signaling, JNK pathway, and calcium signaling. In many cases, these processes or functions have been either implicated in formation of lamellipodia or, conversely, require lamellipodia in order to be carried out in the cell [53-57].

For the Adhesion Disassembly/Cortical Tension Phenoluster, the following gene sets were found to be down-regulated, but at marginal significance (q between 25% and 33 $\frac{1}{3}$%): cytosol, gastrulation, GTPase activity, GEF activity, cell cycle, Wnt signaling, and translation inititation. Several studies have implicated Wnt signaling in the complex regulation of cadherins [58, 59]. The down-regulation of both GTPase and GEF activity is consistent with decreased morphological change associated with adhesion disassembly in this TC group.

For the Rho1 Phenocluster, a number of gene sets were found to be up-regulated, including GO categories for microtubule associated complex and kinesin complex. Additionally, the VEGF pathway and GO categories for GPCR signaling and apoptosis were also up-regulated. For the Rac1 Phenocluster, categories for cytosol and cell cycle regulation are both up-regulated as well. These results are difficult to interpret, but are noted here for completeness. No significant results were obtained for Control versus the Protrusion/Adhesion Formation Phenocluster nor for Control versus other WT-appearing TCs.

Discussion

We elaborated a framework for integrating high-throughput data from morphological and transcriptional genetic screens in order to study genes and pathways involved in different morphological phenotypes. We used previous results in the literature on the genes and pathways that we identified in order to validate our findings and provide support for our framework of study. Overall, our results provided insights into the mechanisms for morphological change.

Statistically significant differential expression of single genes was essentially absent when using standard methods based on t-tests and correction for multiple hypothesis testing. Using a less stringent method (SAM), it was possible to identify single genes exhibiting moderate differential expression for some of the class distinctions under consideration. In many cases, individual genes that were identified by this analysis can be rationalized with the relevant class distinction. Certain patterns of differential expression were evident, and GSEA was performed to rigorously investigate gene set enrichment. Here too, the results of our analyses were in line with known biology, as demonstrated repeatedly above. That the phenocluster comparisons provide biologically valid results supports the use of our methodology for studying variability and its mechanistic explanation using microarray studies.

One must be cautious in interpreting the differential expression results, because the typical average fold change, even for genes found to be significantly differentially expressed, are not as large as one usually finds when performing comparisons of this sort (due to the poor quality of the microarray array data considered here). This highlights the importance of asking the usual question of whether observed changes in expression are biologically significantor not.

As noted earlier, several genes were found to be differentially expressed for the comparison of GFP to other WT-appearing TCs. This finding points to a potential limitation in our approach. Namely, the assignment of TCs to a group for our morphological class distinctions depended on analysis of morphological data, an analysis which is itself subject to error. But an even more fundamental question is raised: must two morphologically indistinguishable cells (or TCs) have precisely the same signaling state? The answer to this question depends on the resolution of our morphological imaging. Since signaling state is ultimately a physical property, we could in theory define appropriate geometrical/morphological traits in order to fully measure signaling state. In practice, however, our morphological imaging is far more coarse than this ideal. Therefore, two cells, or two TCs, identified as physically indistinguishable by our methods, need not be expected to have identical signaling states.

It should be remarked again that the *Drosophila* cell lines used for morphological and transcriptional data were different (BG-2 for morphology, S2R+ for transcription). This may help explain the lack of significant results when using t-tests, along with the relatively poor quality of the microarray dataset under consideration for this analysis, as signal strength is diminished when comparing alternate cell lines. On the other hand, because we do obtain meaningful results even when comparing different cell lines, coupled with the fact that some several previous studies have shown these two cells to share expression/morphology characteristics [60, 61], we are encouraged to carry out further experiments to continue this line of research in future work – namely, to obtain microarray data for a screen using BG-2 cells.

Materials and Methods

Our overarching goal was to study differences in expression between treatment conditions showing different morphologies. This amounted to defining a class distinction to separate treatment conditions into groups of classes on the basis of morphology and then, subsequently, determining differential expression between these groups. We generated two different types of class distinctions corresponding to phenocluster analysis and variability analysis. More specifically, we considered class distinctions defined by: each of six different phenoclusters versus control; and high/low morphological variability versus control and low versus high morphological variability.

Morphological data

As described in [14], TCs were prepared in the *Drosophila* DM-BG2 (referred to as BG-2) cell line using either dsRNA or overexpression constructs. The screen consisted of 249 distinct genetic perturbations, with several replicates, for a total of 273 TCs. The 249 TCs correspond to 45 dsRNAs targeting Rho GTPases, GEFs, and GAPs, 20 overexpression constructs and 173 dsRNAs chosen randomly from a set of genes implicated in cytoskeletal organization, and overexpression of SIF (a *Drosophila* RhoGEF) in combination with several randomly selected dsRNAs. Cell segmentation was performed using the custom CellSegmenter Software. Cells were stochastically labeled with GFP to facilitate image segmentation. For each cell, 145 geometric features and 9 status features were extracted in a semi-automated fashion.

Phenoclusters were used as determined by Bakal et al [14]. Briefly, neural network classifiers were defined as a means of dimensionality reduction from the full 145-dimensional feature space to a reduced, 7-dimensional space. Hierarchical clustering was performed on the set of mean

scores for all TCs. Several clusters were each enriched for genes known to be involved in regulation of a particular morphological process. Here, we used 6 of the 7 phenoclusters identified in this manner (the excluded phenocluster did not have any overlap with the TCs in the transcriptional screen).

Variability of each TC population was determined by Nir et al [21]. Namely, a variability p-score was calculated for each TC in the morphological screen. To accomplish this, raw feature data was first normalized, and then reduced in dimensionality using principal components. The variability v-score for a given TC was defined as a normalized variance of the set of distances of all single cells in the TC from their mean (using the Euclidean distance in reduced feature space). Finally, the variability p-score was determined by bootstrapping to quantify the significance of a variability v-score after accounting for sample size.

Transcriptional data

Microarray data was obtained in the *Drosophila* S2R+ cell line [30] using single-channel Combimatrix custom 4x2k arrays. The screen consisted of 51 distinct genetic perturbations, across a total of 126 TCs (following quality control). Each experiment was characterized by knockout or overexpression of one (typically) or more genes as well as control TCs for each gene chip. Intensity readouts were acquired using Combimatrix software, and logged data was median-centered and loess-normalized. Next, duplicate probes were collapsed to their maximum score, yielding a total of 1832 probes. Finally, missing data was imputed using the nearest-neighbor method. Because the TCs belonging to any of the groups defined by the class

distinctions were well-distributed across the different chips, we did not perform background subtraction to the control TC on each gene chip.

For all subsequent analysis, we restricted our attention to the TCs in the intersection of the sets of TCs for the two screens (**Supplementary Table 1**).

Class distinctions

We defined two main types of class distinctions: phenocluster analysis and variability analysis. We considered class distinctions defined by: each of six different phenoclusters versus control; and high/low morphological variability versus control and low versus high morphological variability (**Supplementary Table 2**). For the variability-based distinctions, we considered the 20 highest- and lowest scoring (variability p-score) TCs in the morphological screen. Those TCs overlapping with the transcriptional screen were the ones used in the final analysis. In each case, the class distinction yielded two groups of TCs. These pairs of groups were subjected to differential expression and gene set enrichment analysis.

Differential expression

Differential expression between two unpaired groups was performed using standard t-tests (Welch t-test for unequal variances, FWER correction at 5% for multiple hypotheses). Separately, we performed significance analysis of microarrays (SAM) to determine differential expression using FDR rates of 5% as a cutoff for significance using 1000 permutations to

calculate FDRs for each comparison. We used the MeV software package for both the t-tests and SAM.

Gene set enrichment

We carried out gene set enrichment analysis (GSEA) for each pair of TC groups, as defined by the class distinction. We made use of the GSEA v 2.0 Java package. All *Drosophila* gene sets in KEGG and GO as of June 2009 were used, as determined by using Flybase precomputed files for *Drosophila* functional annotation [62]. In order to obtain meaningful estimates for FDRs, we required that each gene set under consideration contain at least 15 genes. Given this constraint, the number of gene sets under consideration was reduced to 147 (**Supplementary Tables 3 and 4**). An FDR cutoff of 33 $\frac{1}{3}$% was used, and all other parameters for GSEA were set to their default values.

References

1. Singh R, Xu J, Berger B. Global alignment of multiple protein interaction networks with application to functional orthology detection. Proc. Natl. Acad. Sci. U.S.A. 2008 Sep 2;105(35):12763-12768.

2. Singh R, Xu J, Berger B. Struct2Net: Integrating Structure Into Protein-Protein Interaction Prediction. Pacific Symposium on Biocomputation. 2006.

3. Gat-Viks I, Tanay A, Raijman D, Shamir R. A probabilistic methodology for integrating knowledge and experiments on biological networks. J. Comput. Biol. 2006 Mar ;13(2):165-181.

4. Loots GG, Ovcharenko I. rVISTA 2.0: evolutionary analysis of transcription factor binding sites. Nucl. Acids Res. 2004 Jul 1;32(suppl_2):W217-221.

5. Aerts S, Helden JV, Sand O, Hassan BA. Fine-Tuning Enhancer Models to Predict Transcriptional Targets across Multiple Genomes. PLoS ONE. . 2007 ;2(11):e1115.

6. Markstein M, Markstein P, Markstein V, Levine MS. Genome-wide analysis of clustered Dorsal binding sites identifies putative target genes in the Drosophila embryo. Proc. Natl. Acad. Sci. U.S.A. 2002 Jan 22;99(2):763-768.

7. Hartemink AJ, Segal E. Joint learning from multiple types of genomic data. Pac Symp Biocomput. 2005 ;445-446.

8. Hartemink AJ, Gifford DK, Jaakkola TS, Young RA. Combining location and expression data for principled discovery of genetic regulatory network models. Pac Symp Biocomput. 2002 ;437-449.

9. Djebbari A, Quackenbush J. Seeded Bayesian Networks: constructing genetic networks from microarray data. BMC Syst Biol. 2008 ;257.

10. Mukherjee S, Berger MF, Jona G, Wang XS, Muzzey D, Snyder M, Young RA, Bulyk ML. Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays. Nat. Genet. 2004 Dec ;36(12):1331-1339.

11. Nir et al double knockout

12. Carpenter AE, Jones TR, Lamprecht MR, Clarke C, Kang IH, Friman O, Guertin DA, Chang JH, Lindquist RA, Moffat J, Golland P, Sabatini DM. CellProfiler: image analysis software for identifying and quantifying cell phenotypes. Genome Biol. 2006 ;7(10):R100.

13. Ohya Y, Sese J, Yukawa M, Sano F, Nakatani Y, Saito TL, Saka A, Fukuda T, Ishihara S, Oka S, Suzuki G, Watanabe M, Hirata A, Ohtani M, Sawai H, Fraysse N, Latgé J, François JM, Aebi M, Tanaka S, Muramatsu S, Araki H, Sonoike K, Nogami S, Morishita S. High-dimensional and large-scale phenotyping of yeast mutants. Proc Natl Acad Sci U S A. 2005 Dec 27;102(52):19015-20.

14. Bakal C, Aach J, Church G, Perrimon N. Quantitative Morphological Signatures Define Local Signaling Networks Regulating Cell Morphology. Science. 2007 Jun 22;316(5832):1753-1756.

15. Moffat J, Grueneberg DA, Yang X, Kim SY, Kloepfer AM, Hinkle G, Piqani B, Eisenhaure TM, Luo B, Grenier JK, Carpenter AE, Foo SY, Stewart SA, Stockwell BR, Hacohen N, Hahn WC, Lander ES, Sabatini DM, Root DE. A lentiviral RNAi library for human and mouse genes applied to an arrayed viral high-content screen. Cell. 2006 Mar 24;124(6):1283-98.

16. Glory E, Murphy RF. Automated subcellular location determination and high-throughput microscopy. Dev Cell. 2007 Jan ;12(1):7-16.

17. Perlman ZE, Slack MD, Feng Y, Mitchison TJ, Wu LF, Altschuler SJ. Multidimensional drug profiling by automated microscopy. Science. 2004 Nov 12;306(5699):1194-8.

18. Neumann B, Held M, Liebel U, Erfle H, Rogers P, Pepperkok R, Ellenberg J. High-throughput RNAi screening by time-lapse imaging of live human cells. Nat Methods. 2006 May ;3(5):385-90.

19. Piano F, Schetter AJ, Morton DG, Gunsalus KC, Reinke V, Kim SK, Kemphues KJ. Gene clustering based on RNAi phenotypes of ovary-enriched genes in C. elegans. Curr Biol. 2002 Nov 19;12(22):1959-64.

20. Gil J, Wu H, Wang BY. Image analysis and morphometry in the diagnosis of breast cancer. Microsc Res Tech. 2002 Oct 15;59(2):109-18.

21. Nir et al Variability

22. Levy SF, Siegal ML. Network Hubs Buffer Environmental Variation in Saccharomyces cerevisiae. PLoS Biol. 2008 Nov 4;6(11):e264.

23. Slack MD, Martinez ED, Wu LF, Altschuler SJ. Characterizing heterogeneous cellular responses to perturbations. Proc. Natl. Acad. Sci. U.S.A. 2008 Dec 9;105(49):19306-19311.

24. Spencer SL, Gaudet S, Albeck JG, Burke JM, Sorger PK. Non-genetic origins of cell-to-cell variability in TRAIL-induced apoptosis. Nature. 2009 May 21;459(7245):428-432.

25. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science. 1999 Oct 15;286(5439):531-537.

26. Jiang N, Leach LJ, Hu X, Potokina E, Jia T, Druka A, Waugh R, Kearsey MJ, Luo ZW. Methods for evaluating gene expression from Affymetrix microarray datasets. BMC Bioinformatics. 2008 ;9284.

27. Hatfield GW, Hung S, Baldi P. Differential analysis of DNA microarray gene expression data. Mol. Microbiol. 2003 Feb ;47(4):871-877.

28. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences of the United States of America. 2005 Oct 25;102(43):15545-15550.

29. Mootha VK, Lindgren CM, Eriksson K, Subramanian A, Sihag S, Lehar J, Puigserver P, Carlsson E, Ridderstrale M, Laurila E, Houstis N, Daly MJ, Patterson N, Mesirov JP, Golub TR, Tamayo P, Spiegelman B, Lander ES, Hirschhorn JN, Altshuler D, Groop LC. PGC-1[alpha]-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nat Genet. 2003 Jul ;34(3):267-273.

30. Baym M, Bakal C, Perrimon N, Berger B.: High-Resolution Modeling of Cellular Signaling Networks. Proceedings of the 12th Annual International Conference on Research in Computational Molecular Biology (RECOMB 2008), LNBI 4955: 257-271, 2008

31. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proceedings of the National Academy of Sciences of the United States of America. 2001 Apr 24;98(9):5116-5121.

32. Goshima G, Wollman R, Goodwin SS, Zhang N, Scholey JM, Vale RD, Stuurman N. Genes Required for Mitotic Spindle Assembly in Drosophila S2 Cells. Science. 2007 Apr 20;316(5823):417-421.

33. Björklund M, Taipale M, Varjosalo M, Saharinen J, Lahdenperä J, Taipale J. Identification of pathways regulating cell size and cell-cycle progression by RNAi. Nature. 2006 Feb 23;439(7079):1009-1013.

34. Somma MP, Ceprani F, Bucciarelli E, Naim V, De Arcangelis V, Piergentili R, Palena A, Ciapponi L, Giansanti MG, Pellacani C, Petrucci R, Cenci G, Vernì F, Fasulo B, Goldberg ML, Di Cunto F, Gatti M. Identification of Drosophila mitotic genes by combining co-expression analysis and RNA interference. PLoS Genet. 2008 Jul ;4(7):e1000126.

35. Chen HB, Shen J, Ip YT, Xu L. Identification of phosphatases for Smad in the BMP/DPP pathway. Genes Dev. 2006 Mar 15;20(6):648-653.

36. Moores JN, Roy S, Nicholson DW, Staveley BE. Huntingtin interacting protein 1 can regulate neurogenesis in Drosophila. Eur. J. Neurosci. 2008 Aug ;28(3):599-609.

37. Ye B, Zhang Y, Song W, Younger SH, Jan LY, Jan YN. Growing dendrites and axons differ in their reliance on the secretory pathway. Cell. 2007 Aug 24;130(4):717-729.

38. Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, Itoh M, Katayama T, Kawashima S, Okuda S, Tokimatsu T, Yamanishi Y. KEGG for linking genomes to life and the environment. Nucleic Acids Res. 2008 Jan ;36(Database issue):D480-484.

39. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet. 2000 May ;25(1):25-29.

40. Adam L, Vadlamudi R, Kondapaka SB, Chernoff J, Mendelsohn J, Kumar R. Heregulin regulates cytoskeletal reorganization and cell migration through the p21-activated kinase-1 via phosphatidylinositol-3 kinase. J. Biol. Chem. 1998 Oct 23;273(43):28238-28246.

41. Ma XM, Blenis J. Molecular mechanisms of mTOR-mediated translational control. Nat. Rev. Mol. Cell Biol. 2009 May ;10(5):307-318.

42. Hynes NE, Lane HA. ERBB receptors and cancer: the complexity of targeted inhibitors. Nat. Rev. Cancer. 2005 May ;5(5):341-354.

43. Alroy I, Yarden Y. The ErbB signaling network in embryogenesis and oncogenesis: signal diversification through combinatorial ligand-receptor interactions. FEBS Lett. 1997 Jun 23;410(1):83-86.

44. Avruch J, Hara K, Lin Y, Liu M, Long X, Ortiz-Vega S, Yonezawa K. Insulin and amino-acid regulation of mTOR signaling and kinase activity through the Rheb GTPase. Oncogene. 2006 Oct 16;25(48):6361-6372.

45. Proud CG. Regulation of mammalian translation factors by nutrients. Eur. J. Biochem. 2002 Nov ;269(22):5338-5349.

46. Folkman J. Angiogenesis in cancer, vascular, rheumatoid and other disease. Nat. Med. 1995 Jan ;1(1):27-31.

47. Bianco A, Poukkula M, Cliffe A, Mathieu J, Luque CM, Fulga TA, Rørth P. Two distinct modes of guidance signalling during collective migration of border cells. Nature. 2007 Jul 19;448(7151):362-365.

48. Sims D, Duchek P, Baum B. PDGF/VEGF signaling controls cell size in Drosophila. Genome Biol. 2009 Feb 12;10(2):R20.

49. Perrimon N, Lanjuin A, Arnold C, Noll E. Zygotic lethal mutations with maternal effect phenotypes in Drosophila melanogaster. II. Loci on the second and third chromosomes identified by P-element-induced mutations. Genetics. 1996 Dec ;144(4):1681-1692.

50. Häcker U, Perrimon N. DRhoGEF2 encodes a member of the Dbl family of oncogenes and controls cell shape changes during gastrulation in Drosophila. Genes Dev. 1998 Jan 15;12(2):274-284.

51. Barrett K, Leptin M, Settleman J. The Rho GTPase and a putative RhoGEF mediate a signaling pathway for the cell shape changes in Drosophila gastrulation. Cell. 1997 Dec 26;91(7):905-915.

52. De Calisto J, Araya C, Marchant L, Riaz CF, Mayor R. Essential role of non-canonical Wnt signalling in neural crest migration. Development. 2005 Jun ;132(11):2587-2597.

53. Millard TH, Martin P. Dynamic analysis of filopodial interactions during the zippering phase of Drosophila dorsal closure. Development. 2008 Feb ;135(4):621-626.

54. Martin P, Wood W. Epithelial fusions in the embryo. Curr. Opin. Cell Biol. 2002 Oct ;14(5):569-574.

55. Kim MD, Kolodziej P, Chiba A. Growth cone pathfinding and filopodial dynamics are mediated separately by Cdc42 activation. J. Neurosci. 2002 Mar 1;22(5):1794-1806.

56. Hayden MA, Akong K, Peifer M. Novel roles for APC family members and Wingless/Wnt signaling during Drosophila brain development. Dev. Biol. 2007 May 1;305(1):358-376.

57. Williams MJ, Wiklund M, Wikman S, Hultmark D. Rac1 signalling in the Drosophila larval cellular immune response. J. Cell. Sci. 2006 May 15;119(Pt 10):2015-2024.

58. Jamora C, DasGupta R, Kocieniewski P, Fuchs E. Links between signal transduction, transcription and adhesion in epithelial bud development. Nature. 2003 Mar 20;422(6929):317-322.

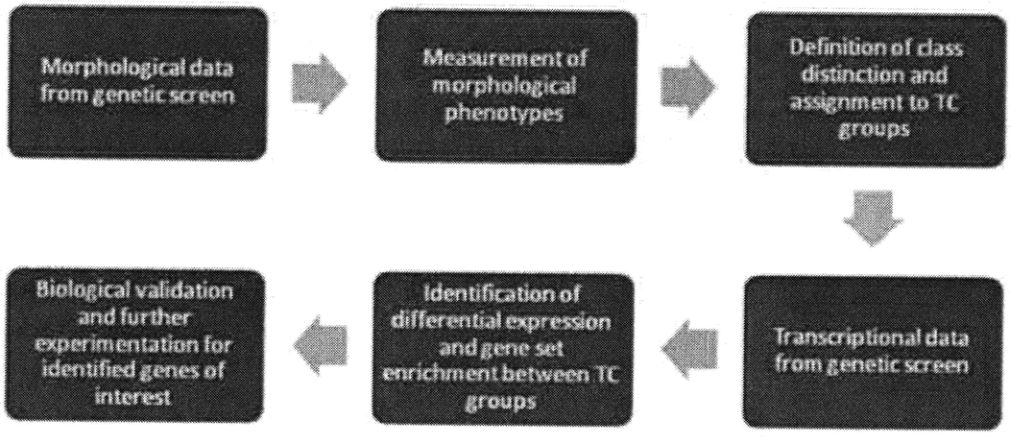
59. Nelson WJ, Nusse R. Convergence of Wnt, beta-catenin, and cadherin pathways. Science. 2004 Mar 5;303(5663):1483-1487.

60. Kiger A, Baum B, Jones S, Jones M, Coulson A, Echeverri C, Perrimon N. A functional genomic analysis of cell morphology using RNA interference. Journal of Biology. 2003 ;2(4):27.

61. Liu T, Sims D, Baum B. Parallel RNAi screens across different cell lines identify generic and cell type-specific regulators of actin organization and cell morphology. Genome Biol. 2009 ;10(3):R26.

62. G. Grumbling, V. Strelets and The FlyBase Consortium (2006). FlyBase: anatomical data, images and queries. *Nucleic Acids Research* **34:** D484-D488; doi:10.1093/nar/gkj068. http://flybase.org/

Figure 1: Workflow for Integration of High-throughput Morphological and Transcriptional Data from Genetic Screens.



Raw morphological data from a genetic screen is first processed using techniques for normalization and dimensionality reduction that have been previously described (top left). Morphological phenotypes are then measured for each of the TCs in the genetic screen (top middle); here we consider phenotypes corresponding to mean TC morphology as well as TC population-level morphological variability. Class distinctions are defined on the basis of the phenotypes that have been measured, and TCs are assigned to one or neither of two groups (top right). Turning to transcriptional data, expression data for these two groups is obtained and preprocessed according to standard techniques (bottom right). Methods for identification of differential expression and gene set enrichment between the two groups are applied (bottom middle). Genes and pathways that have been identified are the subject of further study (bottom left).

Figure 2: Selected SAM plots.

Figure 2A: SAM plot for Low Variability versus High Variability

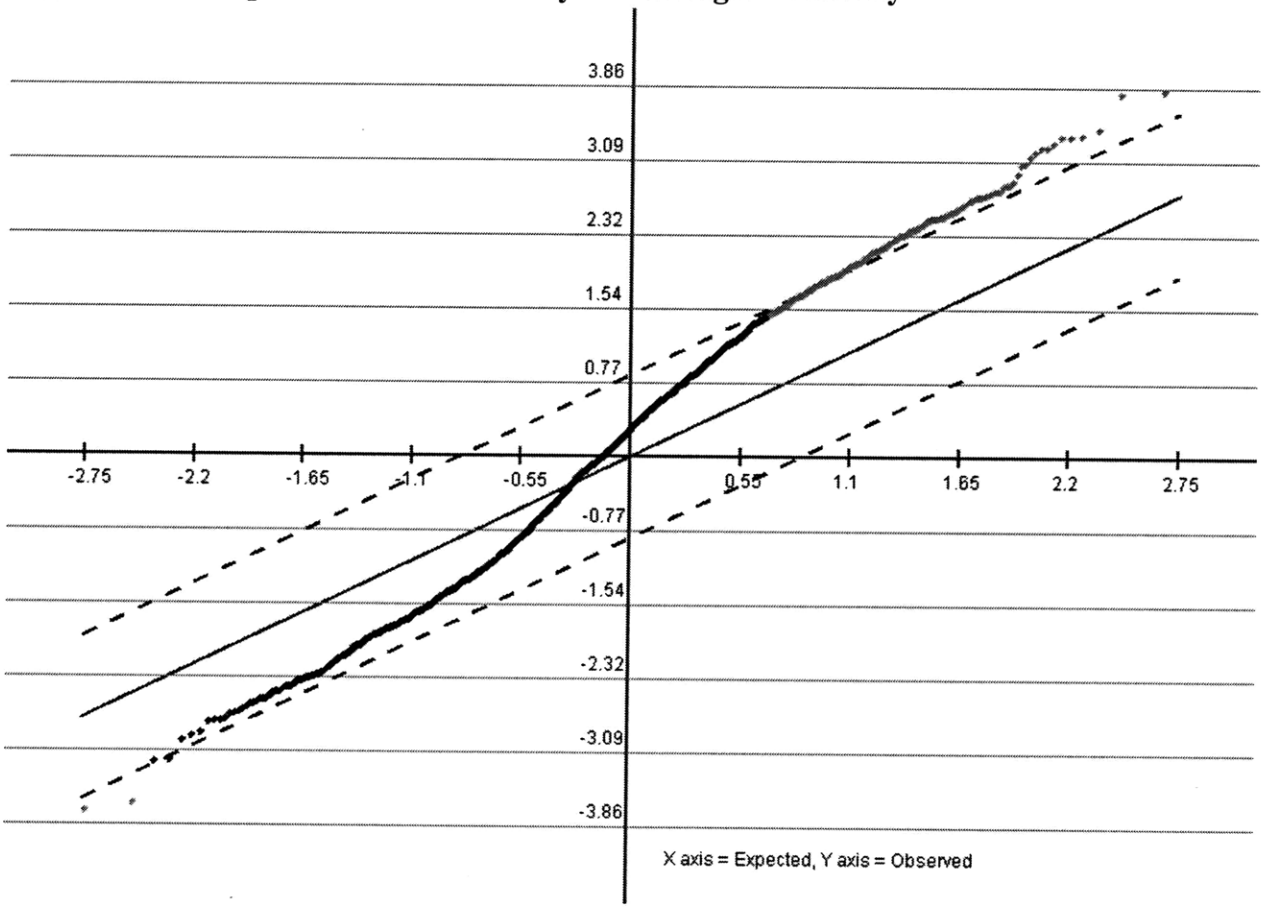


Figure 2B: SAM plot for Control versus the Lamellipodia Formation Phenocluster

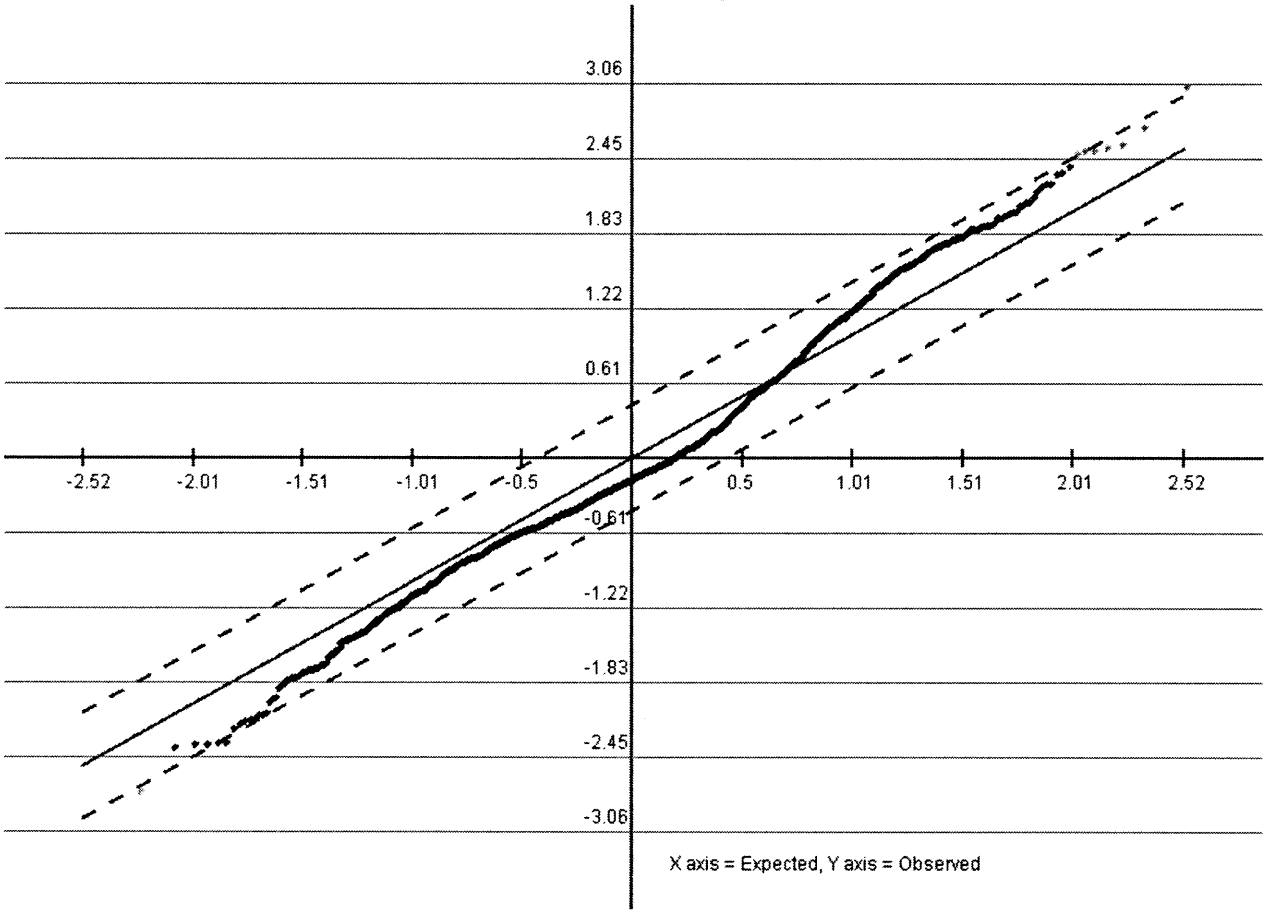
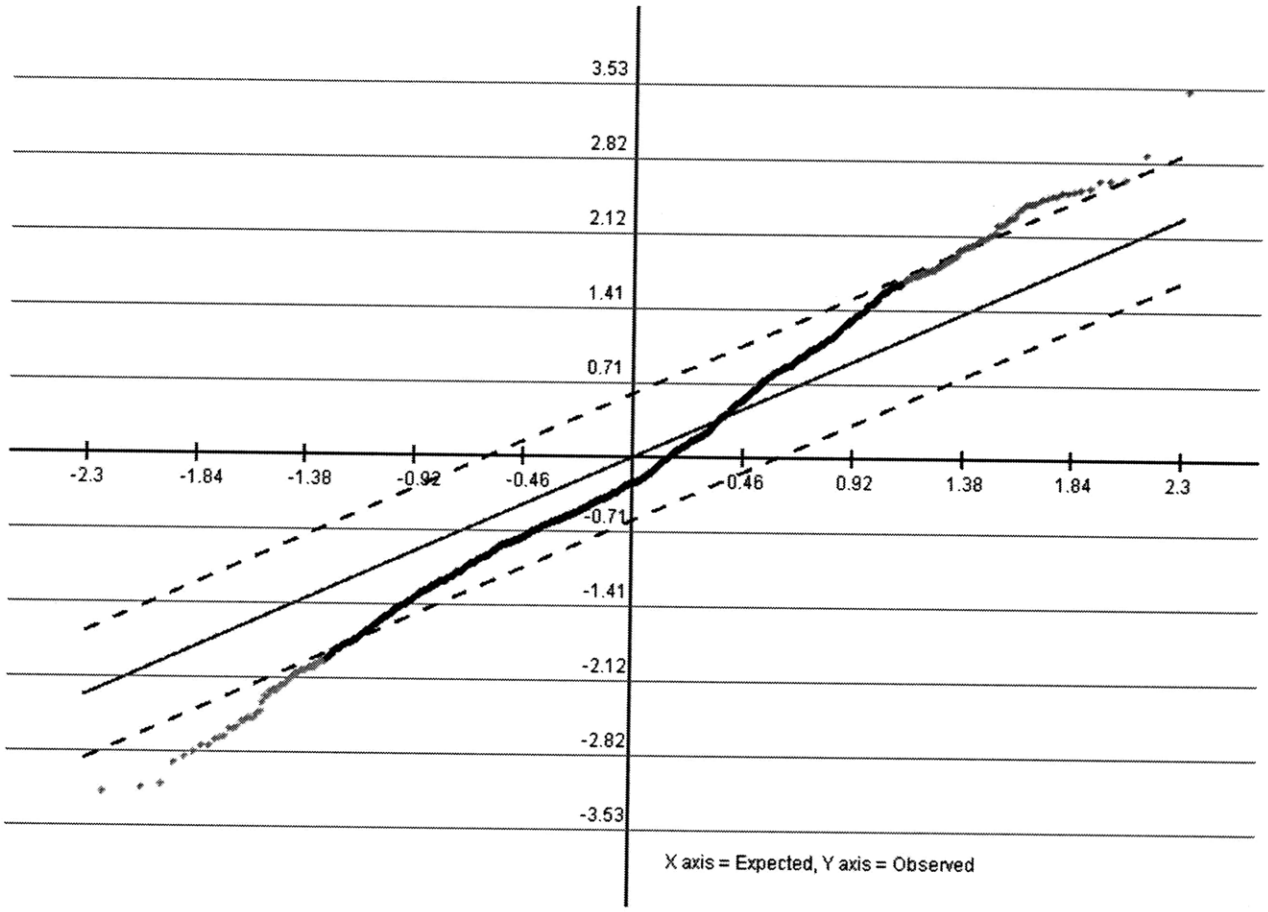


Figure 2C: SAM plot for Control versus the Adhesion Disassembly/Cortical Tension Phenocluster



Selected SAM plots, illustrating results of differential expression analysis. SAM plots are shown for morphological class distinctions for (A) Low Variability versus High Variability, (B) Control versus the Lamellipodia Formation Phenocluster, and (C) Control versus the Adhesion Disassembly/Cortical Tension Phenocluster. The SAM procedure is to compute a normalized coefficient of linear regression for each gene, relative to the class distinction, to determine FDRs for each value using resampling, and finally to identify up- and down-regulated genes by defining an FDR threshold. The SAM plot shows the calculated test statistic for each gene plotted against the expected value, with the genes ordered by expected score. We used FDR significance thresholds of 5% for all comparisons.

Figure 3: Selected GSEA plots for Control versus High Variability.

Figure 3A: ErbB signaling pathway

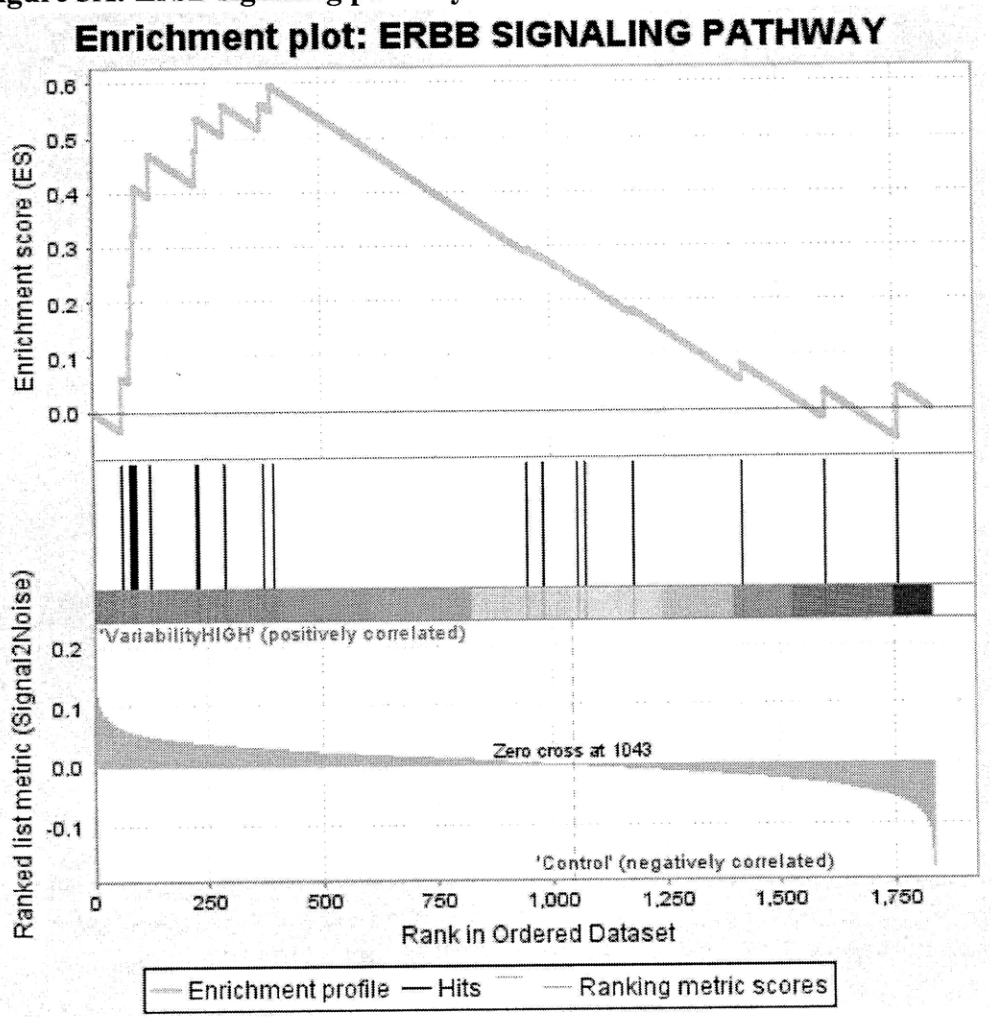
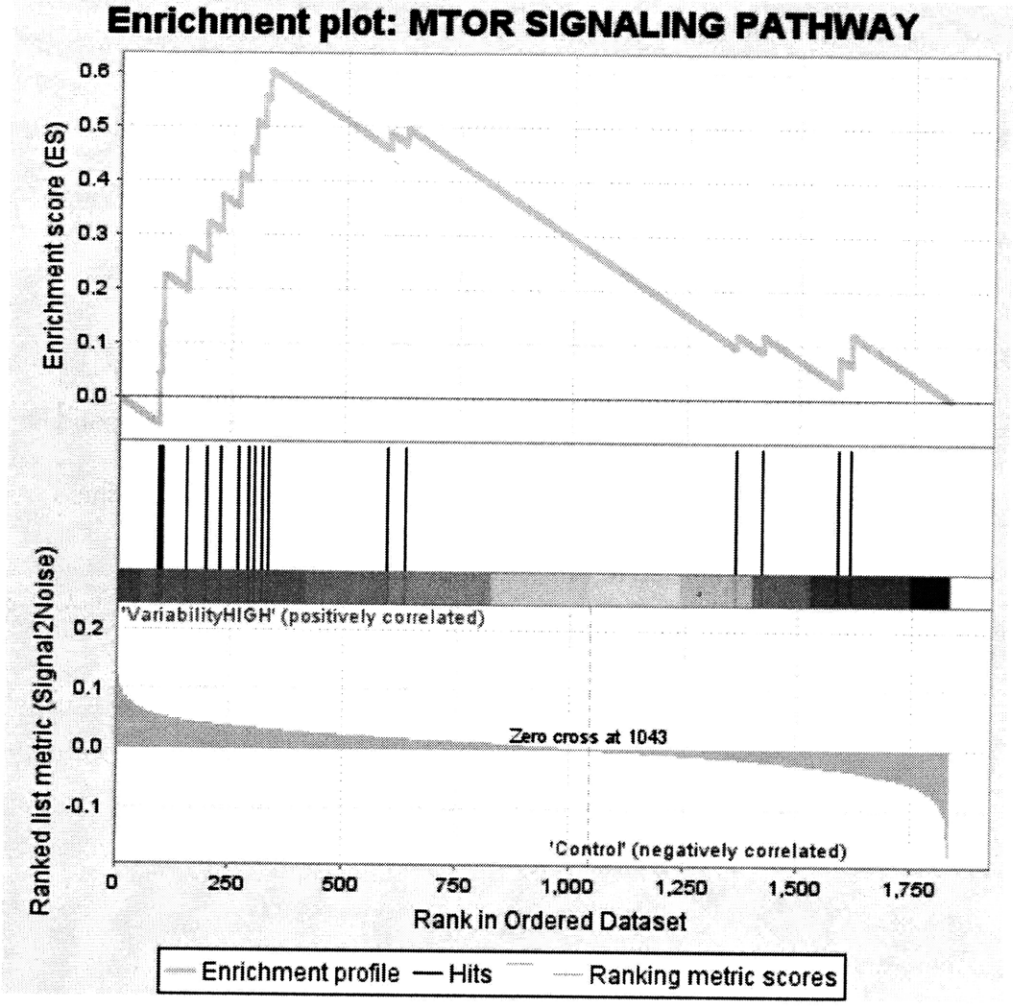


Figure 3B: mTOR signaling pathway



Selected GSEA plots for Control versus High Variability TC groups, illustrating results of gene set enrichment studies. Gene sets for GO categories and KEGG pathways containing at least 15 *Drosophila* genes were analyzed for enrichment among the genes up- or down-regulated, respectively, for the morphological class distinction of Control versus High Variability. The GSEA algorithm proceeds by keeping a running total of a statistic as it traverses the list of gene probes, ordered by correlation with the class distinction (e.g. beginning with the genes most up-regulated in the High Variability group as compared to control). The Enrichment Score (ES) for the gene set is defined to be the maximum statistic encountered in this manner. Subsequently, a

normalized ES (NES) is computed to account for differences in gene sets size and an FDR is computed on the basis of the each NES to account for multiple hypothesis testing.

Two gene sets were identified as being up-regulated in the High Variability TC group, relative to control. Leading edge plots are shown for these gene sets: (A) ErbB signaling pathway and (B) mTor signaling pathway. The x-axis of leading edge plots correspond to genes, ordered by differential expression relative to the class distinction. The vertical black lines indicate the positions of the genes within the gene set under consideration (e.g. ErbB signaling). The solid green line shows the running total of the test statistic as the algorithm progresses through the ordered list of genes. A leading edge plot shifted to one side indicates a greater degree of enrichment than one that peaks near the middle of the gene list.

Figure 4: Selected GSEA plots for Control versus the Lamellipodia Formation Phenocluster.

Figure 4A: Gastrulation

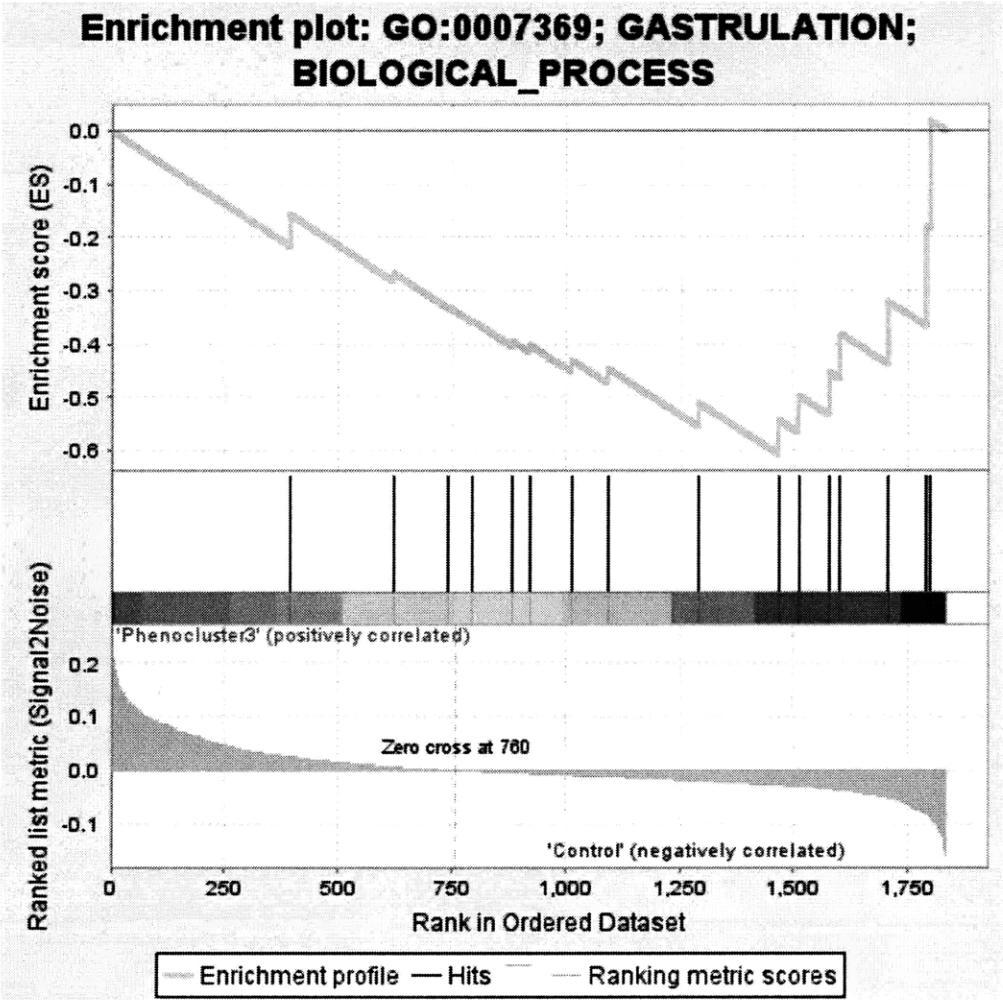


Figure 4B: Cell cycle regulation

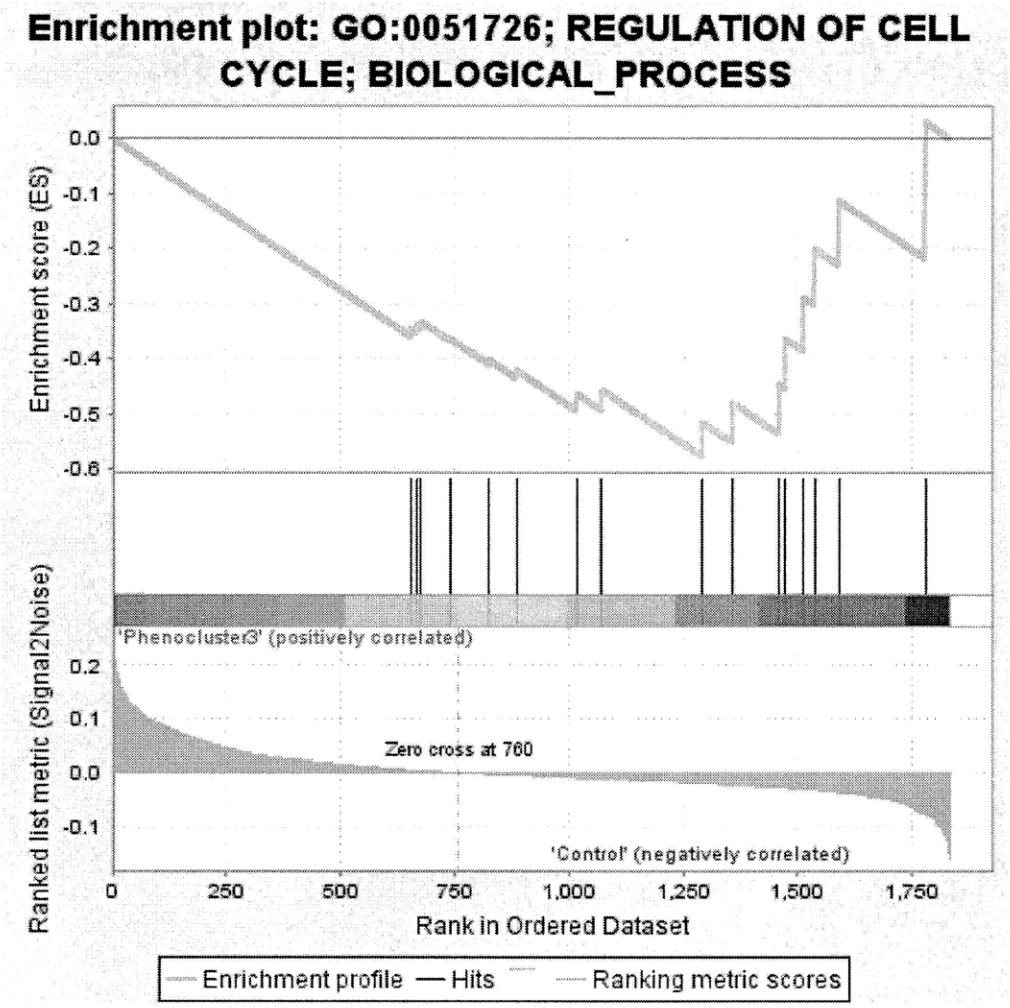


Figure 4C: Wnt signaling pathway

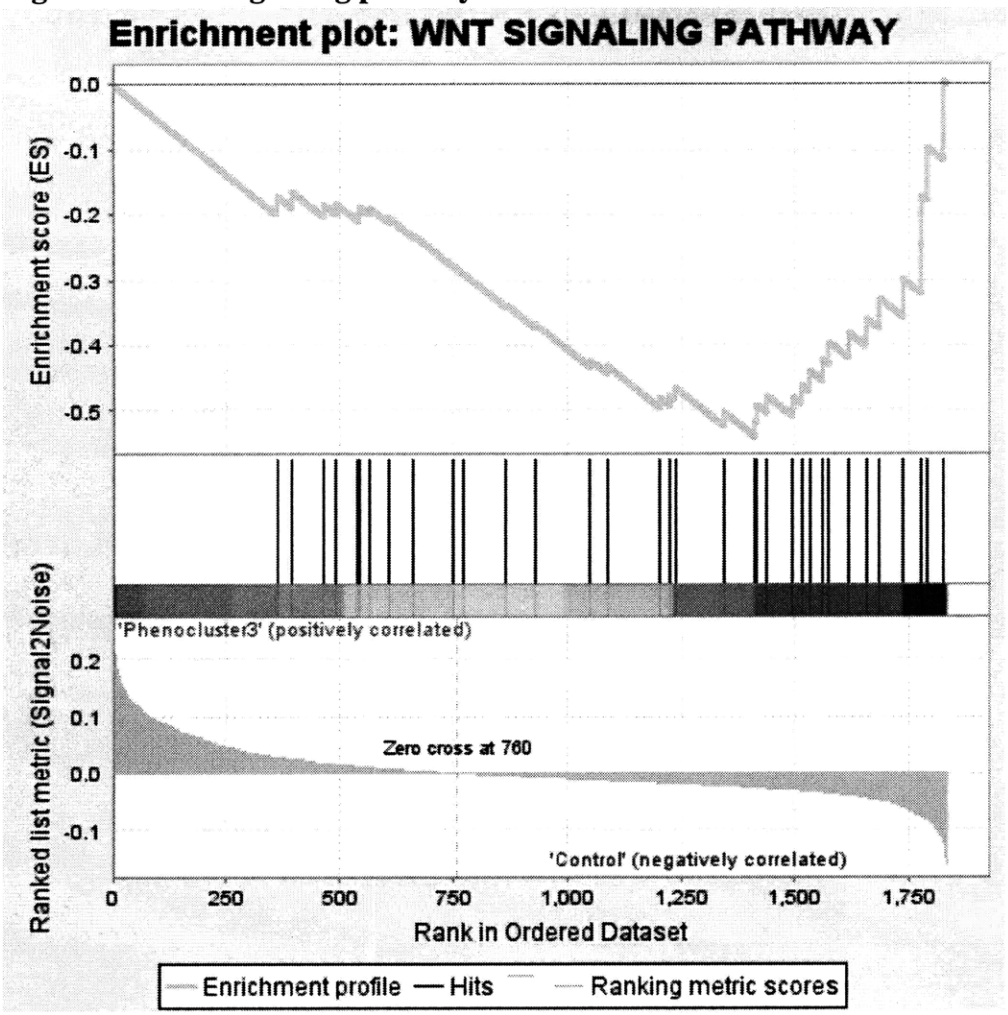
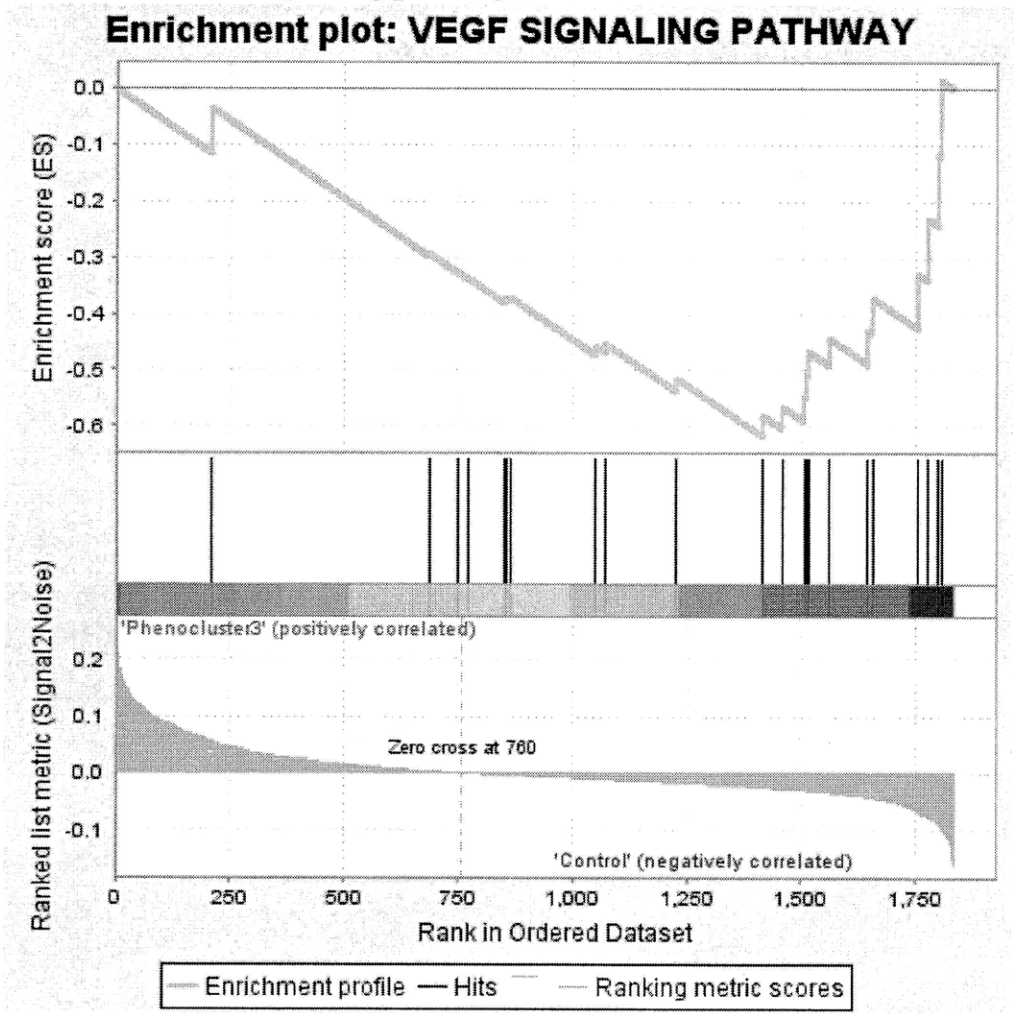


Figure 4D: VEGF signaling pathway



Selected GSEA plots for Control versus the Lamellipodia Formation Phenocluster, illustrating results of gene set enrichment studies. Analogous to **Fig. 3**, GO and KEGG gene sets were subjected to gene set analysis to identify enrichment in up- or down-regulated genes, respectively, for the class distinction for Control versus the Lamellipodia Formation Phenocluster. In total, 29 gene sets were identified as being significantly down-regulated for the Lamellipodia Formation Phenocluster, relative to control, using an FDR threshold of $33\frac{1}{3}\%$. Shown here are leading edge plots for the four gene sets with the lowest FDRs: (A) Gastrulation, (B) Cell cycle regulation, (C) Wnt signaling pathway, and (D) VEGF signaling pathway. See

text for discussion of the significance of these findings, as well as for the other down-regulated gene sets identified by this analysis.

Table 1: Differential expression between TC groups defined by morphological class distinctions.

Class Distinction	Number of Significant Genes (t Tests) (Up / Down)	Number of Significant Genes (SAM) (Up / Down)
Control versus High Variability	0 / 0	2/ 2
Control versus Low Variability	1 / 0	0 / 0
Low versus High Variability	0 / 0	406 / 2
Control versus Rac1 Phenocluster	1 / 0	1 / 2
Control versus Protrusion/Adhesion Formation Phenocluster	0 / 0	0 / 0
Control versus Lamellipodia Formation Phenocluster	0 / 0	7 / 1
Control versus Adhesion Disassembly/Cortical Tension Phenocluster	0 / 0	103 / 61
Control versus GFP/Wild Type Phenocluster	0 / 0	2 / 2
Control versus Rho1 Phenocluster	0 / 0	0 / 1

For each of the class distinctions (see **Supplementary Table 2**), both t tests with FWER multiple hypothesis correction and SAM were applied in order to determine differentially expressed genes between the TC groups defined by the morphological distinction. The number of up and down regulated genes identified by each of the methods is shown here. Note that "up/down" is relative to the second class listed. For example, for the Control vs High Variability distinction, "up-regulated" genes means those that are up-regulated in the High Variability TC group. For the t tests, the lone significant genes were CG4041, which is a Rab GAP (for Control versus Low Variability), and CG9619, which is a phosphatase with unknown biological role (for Control versus the Rac1 Phenocluster). See text for further discussion.

Table 2: Results of SAM analysis.

Table 2A: Control vs High Variability

Gene	FC	FDR (%)	GO Annotation
CG30440	1.370246	0	guanyl-nucleotide exchange factor activity; intracellular; Rho guanyl-nucleotide exchange factor activity; regulation of Rho protein signal transduction
Rel	1.180226	0	transcription factor activity; specific RNA polymerase II transcription factor activity; protein binding; nucleus; cytoplasm; regulation of transcription, DNA-dependent; immune response; positive regulation of antibacterial peptide biosynthetic process; positive regulation of antifungal peptide biosynthetic process; signal transduction; Toll signaling pathway; response to bacterium; antimicrobial humoral response; cellular response to amino acid starvation; innate immune response; positive regulation of innate immune response; regulation of transcription; positive regulation of transcription from RNA polymerase II promoter; defense response to Gram-negative bacterium
CG11583	0.82596	0	molecular_function; nucleolus; ribosomal large subunit biogenesis
bcn92	0.810263	0	[Unknown function]

(Lack of shading indicates up-regulation in the second category – "High Variability" in this case, while shading indicates up-regulation in the first category – "Control" in this case.)

Table 2B: Control vs Low Variability

[No significant results using SAM]

Table 2C: Low vs High Variability

Gene	FC	FDR (%)	GO Annotation
Ef2b	1.642785	0.99189	mitotic spindle elongation; translation elongation factor activity; GTPase activity; GTP binding; cytoplasm; lipid particle; cytosol; translation; translational elongation; mitotic spindle organization
fok	1.620268	0	[Unknown function]
Gp150 (FBgn0013272)	1.579235	0	catalytic activity; protein binding; ATP binding; plasma membrane; transmembrane receptor protein tyrosine phosphatase signaling pathway; metabolic process; compound eye development
Tao-1	1.516619	0.467558	protein serine/threonine kinase activity; receptor signaling

(FBgn0031030)			protein serine/threonine kinase activity; ATP binding; protein amino acid phosphorylation; apoptosis
Protein tyrosine phosphatase 4E (FBgn0004368)	1.505012	0	protein tyrosine phosphatase activity; transmembrane receptor protein tyrosine phosphatase activity; plasma membrane; protein amino acid dephosphorylation
wibg	0.662613	0	[Unknown function]
CG3891	0.574558	0	transcription factor activity; nucleus; regulation of transcription, DNA-dependent; phagocytosis, engulfment

Table 2D: Control vs Rac1 Phenocluster

Gene	FC	FDR (%)	GO Annotation
CG4853	1.202132	0	deoxyribonuclease activity; Ras guanyl-nucleotide exchange factor activity; intracellular; DNA repair; DNA recombination; mushroom body development; regulation of small GTPase mediated signal transduction
tiptop	0.754101	0	DNA binding; transcription factor activity; nucleus; regulation of transcription, DNA-dependent; specification of segmental identity, head; zinc ion binding; regulation of transcription; epidermis morphogenesis
Rac1	0.624666	0	phagocytosis triggered by activation of immune response cell surface activating receptor; GTPase activity; GTP binding; intracellular; phagocytosis, engulfment; microtubule-based process; establishment of tissue polarity; JNK cascade; small GTPase mediated signal transduction; border follicle cell migration; germ-band shortening; dorsal closure; dorsal closure, elongation of leading edge cells; axonogenesis; axon guidance; axonal fasciculation; ventral cord development; peripheral nervous system development; open tracheal system development; tracheal outgrowth, open tracheal system; salivary gland morphogenesis; hemocyte development; myoblast fusion; head involution; cell proliferation; glial cell migration; rhabdomere; morphogenesis of larval imaginal disc epithelium; lamellipodium assembly; actin cytoskeleton organization; ovarian follicle cell development; adherens junction maintenance; hemocyte migration; cell competition in a multicellular organism; imaginal disc-derived wing hair site selection; rhabdomere development; establishment of ommatidial polarity; cell-cell junction organization; regulation of hemocyte differentiation; dorsal closure, amnioserosa morphology change; dorsal appendage formation; axon extension; muscle fiber development; dendrite morphogenesis; regulation of dendrite morphogenesis; regulation of axonogenesis; regulation of synapse organization; actin filament bundle formation; myoblast proliferation

Table 2E: Control vs Protrusion/Adhesion Formation Phenocluster
[No significant results using SAM]

Table 2F: Control vs Lamellipodia Formation Phenocluster

Gene	FC	FDR (%)	GO Annotation
CG30320	1.661964	0	[Unknown function]
CG11597 (FBgn0036212)	1.555574	0	protein phosphatase type 2A complex; protein serine/threonine phosphatase activity; protein amino acid dephosphorylation
CG31157	1.511413	0	[Unknown function]
Rab40 (CG1900)	1.463111	0	GTPase activity; GTP binding; small GTPase mediated signal transduction; regulation of cell shape; protein transport
CG8641	1.410649	0	GTPase activity; GTP binding; intracellular; small GTPase mediated signal transduction
CG8636	0.756723	0	mitotic spindle elongation; nucleotide binding; nucleic acid binding; mRNA binding; translation initiation factor activity; eukaryotic translation initiation factor 3 complex; translational initiation; mitotic spindle organization; zinc ion binding

Table 2G: Control vs Adhesion Disassembly/Cortical Tension Phenocluster

Gene	FC	FDR (%)	GO Annotation
CG30320	1.472293	0	[Unknown function]
Huntingtin interacting protein 1 (CG10971)	1.346096	3.789641	actin binding; phospholipid binding; cytoskeleton organization
CG11597 (FBgn0036212)	1.341283	0	protein phosphatase type 2A complex; protein serine/threonine phosphatase activity; protein amino acid dephosphorylation
Sap47	1.330561	1.405277	[Unknown function]
CG14644	1.320278	0	[Unknown function]
Sar1 (CG7073)	0.777885	0	GTPase activity; GTP binding; lipid particle; intracellular protein transport; larval chitin-based cuticle development; ER to Golgi transport vesicle; embryonic heart tube development; chitin-based larval cuticle pattern formation; negative regulation of dendrite morphogenesis
Ef1gamma	0.773094	1.435827	translation elongation factor activity; lipid particle; cytosol; eukaryotic translation elongation factor 1 complex; translational elongation; salivary gland cell autophagic cell death; autophagic cell death
eIF-5C	0.770507	0	protein binding; cytoplasm; long-term memory; axon

			midline choice point recognition; negative regulation of translation; axon; translation initiation factor binding; ribosome binding; cell soma; oogenesis; neuron fate commitment
Trip1	0.766404	0	translation initiation factor activity; cytosol; eukaryotic translation initiation factor 3 complex; translation; translational initiation
CG16817	0.741179	0	[Unknown function]

Table 2H: Control vs GFP/Wild Type Phenocluster

Gene	FC	FDR (%)	GO Annotation
Dredd	1.132963	0	cysteine-type endopeptidase activity; protein binding; cytoplasm; proteolysis; apoptosis; defense response; immune response; positive regulation of antibacterial peptide biosynthetic process; sperm individualization; apoptotic protease activator activity; innate immune response; defense response to Gram-negative bacterium
CG30065	1.115226	0	[Unknown function]
CG9426	0.887825	0	actin binding; protein binding; cytoplasm; cytoplasmic sequestering of transcription factor
CG12022	0.704688	0	[Unknown function]

Table 2I: Control vs Rho1 Phenocluster

Gene	FC	FDR (%)	GO Annotation
Rho1	0.827193	0	cytokinesis; establishment of planar polarity; establishment of imaginal disc-derived wing hair orientation; GTPase activity; protein binding; GTP binding; cell cortex; endocytosis; cytoskeleton organization; actin filament organization; establishment of tissue polarity; epidermal growth factor receptor signaling pathway; JNK cascade; small GTPase mediated signal transduction; cellularization; determination of left/right symmetry; gastrulation; ventral furrow formation; posterior midgut invagination; germ-band extension; dorsal closure; dorsal closure, spreading of leading edge cells; neuroblast proliferation; axon guidance; peripheral nervous system development; open tracheal system development; motor axon guidance; glial cell migration; germ cell migration; gastrulation involving germ band extension; Wnt receptor signaling pathway; muscle attachment; ommatidial rotation; kinase binding; actin cytoskeleton organization; myofibril assembly; hemocyte migration; branch fusion, open tracheal system; lumen formation, open tracheal system; spiracle morphogenesis, open tracheal system; regulation of Malpighian tubule

			size; imaginal disc-derived wing hair organization; wound healing; establishment of protein localization; dorsal closure, leading edge cell differentiation; dorsal closure, amnioserosa morphology change; dendrite morphogenesis; regulation of axonogenesis; actin filament bundle formation

Selected results of SAM analysis. For each of the nine class distinctions, SAM was performed with 1000 permutations and an FDR cutoff of 5% to determine differentially expressed genes. These genes were then ranked by fold change (FC) with the highest and lowest five genes listed here (unless there were fewer than five at either extreme– see **Table 1**). Note that there were no significant genes in the case of Control versus Low Variability and Control versus the Protrusion/Adhesion Formation Phenocluster (though there was a significant gene in the former case by t tests). Rac1 and Rho1 are identified as being downregulated in the Rac1 Phenocluster and Rho1 Phenocluster, respectively, relative to control. The largest number of differentially expressed genes arose from consideration of the Low versus High Variability and Control versus Adhesion Disassembly/Cortical Tension class distinctions. The fact that so many genes were differentially expressed in the High Variability group (versus the Low Variability group) may be significant biologically, namely it may be a reflection of the abnormal expression and signaling states that distinguish high and low variability populations. Further experimentation is necessary to investigate this hypothesis. The gene CG30320 of unknown function as well as the serine/threonine phosphatase CG11597 are up-regulated for both the Control versus Lamellipodia Formation Phenocluster and the Control versus Adhesion Disassembly/Cortical Tension Phenocluster class distinctions. See also the main text and **Fig. 2**.

Table 3: Results of GSEA analysis.

Table 3A: Control vs High Variability

NAME	SIZE	ES	NES	NOM p-val	FDR q-val	FWER p-val
ERBB SIGNALING PATHWAY	19	0.59401	1.570552	0.034951	0.299927	0.56
MTOR SIGNALING PATHWAY	17	0.60207	1.630138	0.011834	0.324667	0.375

Table 3B: Control vs Low Variability

NAME	SIZE	ES	NES	NOM p-val	FDR q-val	FWER p-val
GO:0007369; GASTRULATION; BIOLOGICAL_PROCESS	16	-0.51629	-1.61065	0.019342	0.203589	0.397
VEGF SIGNALING PATHWAY	22	-0.58382	-1.70696	0.021154	0.211836	0.188
GO:0051726; REGULATION OF CELL CYCLE; BIOLOGICAL_PROCESS	16	-0.4907	-1.61323	0.01833	0.298905	0.392

(Lack of shading indicates up-regulation in the second category, while shading indicates up-regulation in the first category.)

Table 3C: Low vs High Variability

[No significant results using GSEA]

Table 3D: Control vs Rac1 Phenocluster

NAME	SIZE	ES	NES	NOM p-val	FDR q-val	FWER p-val
GO:0005829; CYTOSOL; CELLULAR_COMPONENT	44	0.560872	1.583478	0.029038	0.267824	0.508
GO:0051726; REGULATION OF CELL CYCLE; BIOLOGICAL_PROCESS	16	0.505506	1.620601	0.025048	0.280078	0.407

Table 3E: Control vs Protrusion/Adhesion Formation Phenocluster
[No significant results using GSEA]

Table 3F: Control vs Lamellipodia Formation Phenocluster

NAME	SIZE	ES	NES	NOM p-val	FDR q-val	FWER p-val
GO:0007369; GASTRULATION; BIOLOGICAL_PROCESS	16	-0.60712	-1.8544	0.004228	0.022635	0.042
GO:0051726; REGULATION OF CELL CYCLE; BIOLOGICAL_PROCESS	16	-0.57677	-1.90521	0	0.02333	0.022
WNT SIGNALING PATHWAY	35	-0.54063	-1.77295	0.002096	0.030064	0.092
VEGF SIGNALING PATHWAY	22	-0.61882	-1.77625	0.016227	0.038501	0.088
GO:0030036; ACTIN CYTOSKELETON ORGANIZATION; BIOLOGICAL_PROCESS	22	-0.6449	-1.57775	0.016427	0.09423	0.473
GO:0007409; AXONOGENESIS; BIOLOGICAL_PROCESS	16	-0.56799	-1.5844	0.029536	0.09718	0.454
GO:0006413; TRANSLATIONAL INITIATION; BIOLOGICAL_PROCESS	41	-0.61401	-1.5636	0.028986	0.098835	0.519
GO:0005198; STRUCTURAL MOLECULE ACTIVITY; MOLECULAR_FUNCTION	26	-0.49844	-1.59093	0.006224	0.100906	0.434
GO:0005834; HETEROTRIMERIC G-PROTEIN COMPLEX; CELLULAR_COMPONENT	19	-0.59238	-1.60124	0.032389	0.10237	0.407
GO:0007391; DORSAL CLOSURE; BIOLOGICAL_PROCESS	49	-0.50074	-1.61845	0.025948	0.102998	0.37
GO:0003743; TRANSLATION INITIATION FACTOR	45	-0.63274	-1.6258	0.018634	0.11171	0.356

ACTIVITY; MOLECULAR_FUNCTION						
GO:0042803; PROTEIN HOMODIMERIZATION ACTIVITY; MOLECULAR_FUNCTION	16	-0.6561	-1.53418	0.0409	0.112665	0.602
CALCIUM SIGNALING PATHWAY	24	-0.50859	-1.53855	0.040984	0.11607	0.597
GO:0008258; HEAD INVOLUTION; BIOLOGICAL_PROCESS	16	-0.63876	-1.52091	0.036072	0.118769	0.636
GO:0005811; LIPID PARTICLE; CELLULAR_COMPONENT	27	-0.68857	-1.51201	0.014799	0.120205	0.664
GO:0005829; CYTOSOL; CELLULAR_COMPONENT	44	-0.65278	-1.63557	0.021053	0.120574	0.333
GO:0005938; CELL CORTEX; CELLULAR_COMPONENT	18	-0.48882	-1.47342	0.060417	0.156269	0.758
GO:0003924; GTPASE ACTIVITY; MOLECULAR_FUNCTION	111	-0.39347	-1.38701	0.095238	0.239584	0.92
GO:0006412; TRANSLATION; BIOLOGICAL_PROCESS	20	-0.59667	-1.39918	0.092213	0.242605	0.902
GO:0005525; GTP BINDING; MOLECULAR_FUNCTION	115	-0.3943	-1.39153	0.08026	0.243434	0.911
GO:0048477; OOGENESIS; BIOLOGICAL_PROCESS	55	-0.40291	-1.40402	0.033597	0.247148	0.893
GO:0007254; JNK CASCADE; BIOLOGICAL_PROCESS	19	-0.52445	-1.36838	0.113684	0.24829	0.941
GO:0006911; PHAGOCYTOSIS, ENGULFMENT; BIOLOGICAL_PROCESS	45	-0.4946	-1.37281	0.116525	0.251373	0.938
GO:0005516; CALMODULIN BINDING; MOLECULAR_FUNCTION	16	-0.48651	-1.35704	0.122699	0.257189	0.953
MTOR SIGNALING PATHWAY	17	-0.49414	-1.32498	0.170124	0.307748	0.971
GO:0007269; NEUROTRANSMITTER SECRETION; BIOLOGICAL_PROCESS	18	-0.4444	-1.31745	0.11002	0.309759	0.974

GO:0006915; APOPTOSIS; BIOLOGICAL_PROCESS	16	-0.51234	-1.30144	0.170168	0.31532	0.98
GO:0007186; G-PROTEIN COUPLED RECEPTOR PROTEIN SIGNALING PATHWAY; BIOLOGICAL_PROCESS	29	-0.36231	-1.29227	0.180361	0.32022	0.982
GO:0007507; HEART DEVELOPMENT; BIOLOGICAL_PROCESS	24	-0.36199	-1.30241	0.109434	0.325474	0.978

Table 3G: Control vs Adhesion Disassembly/Cortical Tension Phenocluster

NAME	SIZE	ES	NES	NOM p-val	FDR q-val	FWER p-val
GO:0005829; CYTOSOL; CELLULAR_COMPONENT	44	-0.62355	-1.57432	0.026585	0.258879	0.559
GO:0007369; GASTRULATION; BIOLOGICAL_PROCESS	16	-0.47547	-1.48353	0.049336	0.285343	0.802
GO:0003924; GTPASE ACTIVITY; MOLECULAR_FUNCTION	111	-0.412	-1.46187	0.026694	0.304508	0.836
GO:0003743; TRANSLATION INITIATION FACTOR ACTIVITY; MOLECULAR_FUNCTION	45	-0.57658	-1.48597	0.07551	0.311022	0.797
GO:0005085; GUANYL-NUCLEOTIDE EXCHANGE FACTOR ACTIVITY; MOLECULAR_FUNCTION	21	-0.50282	-1.44404	0.068226	0.319147	0.868
WNT SIGNALING PATHWAY	35	-0.47969	-1.52938	0.035503	0.321172	0.693
GO:0051726; REGULATION OF CELL CYCLE; BIOLOGICAL_PROCESS	16	-0.47806	-1.5747	0.041825	0.322273	0.557

Table 3H: Control vs GFP/Wild Type Phenocluster

[No significant results using GSEA]

Table 3I: Control vs Rho1 Phenocluster

NAME	SIZE	ES	NES	NOM p-val	FDR q-val	FWER p-val
VEGF SIGNALING PATHWAY	22	0.621663	1.788412	0.00813	0.030902	0.096
GO:0007186; G-PROTEIN COUPLED RECEPTOR PROTEIN SIGNALING PATHWAY; BIOLOGICAL_PROCESS	29	0.49755	1.807913	0.004246	0.031376	0.075
GO:0005834; HETEROTRIMERIC G-PROTEIN COMPLEX; CELLULAR_COMPONENT	19	0.651063	1.829237	0.006048	0.033802	0.057
GO:0006350; TRANSCRIPTION; BIOLOGICAL_PROCESS	15	0.560483	1.752683	0.009579	0.038484	0.134
GO:0004871; SIGNAL TRANSDUCER ACTIVITY; MOLECULAR_FUNCTION	32	0.549064	1.870642	0	0.041301	0.036
GO:0005871; KINESIN COMPLEX; CELLULAR_COMPONENT	24	0.555139	1.719027	0.002053	0.049436	0.19
GO:0005875; MICROTUBULE ASSOCIATED COMPLEX; CELLULAR_COMPONENT	47	0.439442	1.665025	0	0.074844	0.302
GO:0006915; APOPTOSIS; BIOLOGICAL_PROCESS	16	0.643057	1.638487	0.010081	0.083756	0.362
CALCIUM SIGNALING PATHWAY	24	0.524027	1.60669	0.01636	0.101751	0.444
GO:0035071; SALIVARY GLAND CELL AUTOPHAGIC CELL DEATH; BIOLOGICAL_PROCESS	26	0.576788	1.537963	0.025194	0.144126	0.64
GO:0007067; MITOSIS; BIOLOGICAL_PROCESS	24	0.555802	1.545188	0.008403	0.147451	0.617
GO:0007155; CELL ADHESION; BIOLOGICAL_PROCESS	37	0.480576	1.523833	0.01222	0.153675	0.696
GO:0048102; AUTOPHAGIC CELL DEATH; BIOLOGICAL_PROCESS	24	0.597827	1.550784	0.02924	0.154886	0.603
GO:0005089; RHO	21	0.497489	1.38469	0.09407	0.320404	0.952

GUANYL-NUCLEOTIDE EXCHANGE FACTOR ACTIVITY; MOLECULAR_FUNCTION						
GO:0007420; BRAIN DEVELOPMENT; BIOLOGICAL_PROCESS	18	0.481422	1.371221	0.098196	0.331259	0.969

Results of GSEA analysis. GO annotations and KEGG pathways with at least fifteen genes in the common with the microarray probe/gene set were used (see **Supplementary Tables 3** and **4** for a complete list of the 147 gene sets used for GSEA). Analysis was run using 1000 permutations and an FDR cutoff of 33 $\frac{1}{3}$% to determine significance. Pathways/GO categories are up-regulated for the Rac1 and Rho1 Phenoclusters and for the High Variability group, while they are down-regulated for the Lamellipodia Formation Phenocluster, the Adhesion Disassembly/Cortical Tension Phenocluster, and the Low Variability group (as indicated by appropriate shading in the tables above) No significant results were observed for the Protrusion/Adhesion Formation Phenocluster, the GFP/Wild-type Phenocluster, nor for the Low versus High Variability comparison. It was expected that the GFP/Wild-type Phenocluster should not differ significantly from Control, since the morphological distinction in that case is effectively absent; this serves as a sort of "negative control" result. See main text for extensive discussion as well as **Fig. 3**.

Data integration for high-throughput morphological and transcriptional genetic screens

Supplementary figures and text:

Supplementary Table 1	TCs included in both the morphological and transcriptional screens
Supplementary Table 2	Morphological class distinctions
Supplementary Table 3	KEGG pathways for GSEA
Supplementary Table 4	GO categories for GSEA

Supplementary Table 1

TCs included in both the morphological and transcriptional screens

Treatment Condition	# replicates	Phenocluster	High/low variability?
CG10188	2		
CG14045	1		
CG15611	1	C	
CG30115	1	E	
CG30456	1	D	High
CG3799	4	E	
CG3799.over	1	E	
CdGAPr	2	F	
Cdc42	4	E	High
Cdc42Y32A	1	E	
Cdep	1		High
GFP	29	Control	Control
MTL	2		
Rac1	6	A	
RacF28L	2	D	Low
RacGAP50C	4	D	
RacV12	1	D	Low
Rho1	5	F	
RhoBTB	4		
RhoF30L	2	C	Low
RhoGAP100F	2	D	
RhoGAP16F	1	F	
RhoGAP19D	1	D	
RhoGAP1A	4	F	
RhoGAP54D	2	D	
RhoGAP5A	1	F	
RhoGAP92B	1	E	
RhoGAP93B	2	F	
RhoGAPp190	2	B	
RhoGEF2	1	F	
RhoGEF3	1	B	
RhoGEF4	2	F	
RhoGEF64C	2	C	
RhoGEF64C.over	1	A	
RhoL	4	B	

RhoV14	1	E	
oncoCG3799.over	1		
oncoGEF3.over	1		
oncoSif.over	2		
pbl	3	D	High
sif	4		
sif.over	1		
sifFL.over	1	D	
trio	1	C	

Supplementary Table 1. TCs included in both the morphological and transcriptional screens. We used data from a 273-TC (including replicates) morphological screen and a 126-TC (including replicates) transcriptional screen. A total of 114 TCs(including replicates) of the transcriptional screen had analogues in the morphological screen, representing 44 unique TCs. The unique TCs are listed here, with number of chip replicates indicated in the second column. The third column indicates to which of the six phenoclusters, if any, the TC belongs (legend: A – Rac1 Phenocluster; B – Protrusion/Adhesion Formation Phenocluster; C – Lamellipodia Formation Phenocluster; D – Adhesion Disassembly/Cortical Tension Phenocluster; E – GFP/Wild Type Phenocluster; F –Rho1 Phenocluster). The fourth column indicates whether the TC had significantly high or low morphological variability.

Supplementary Table 2
Morphological class distinctions

Class Distinction	# TCs in First Group	# TCs in Second Group
Control versus High Variability	29	9
Control versus Low Variability	29	5
Low versus High Variability	5	9
Control versus Rac1 Phenocluster	29	7
Control versus Protrusion/Adhesion Formation Phenocluster	29	6
Control versus Lamellipodia Formation Phenocluster	29	6
Control versus Adhesion Disassembly/Cortical Tension Phenocluster	29	17
Control versus GFP/Wild Type Phenocluster	29	13
Control versus Rho1 Phenocluster	29	18

Supplementary Table 2. Morphological class distinctions. We considered nine morphological class distinctions, as listed in the leftmost column. The second and third columns indicate the number of TCs (counting replicates) in each of the two TC groups defined by the class distinctions.

Supplementary Table 3
KEGG pathways for GSEA

KEGG Pathway
Hedgehog signaling pathway
mTor signaling pathway
ErbB signaling pathway
VEGFsignaling pathway
Wnt signaling pathway
Glycerolipid metabolism
Glycerophospholipid metabolism
Phosphatidylinositol signaling system
Inositol phosphate metabolism
Calcium signaling pathway
Purine metabolism

Supplementary Table 3. KEGG pathways for GSEA. A minimum of fifteen genes was required for inclusion as a gene set for GSEA. The 11 KEGG pathways containing at least fifteen *Drosophila* genes are listed here. See text for additional discussion of input data and parameters for GSEA.

Supplementary Table 4
GO categories for GSEA

GO Annotation
GO:0000122; NEGATIVE REGULATION OF TRANSCRIPTION FROM RNA POLYMERASE II PROMOTER; BIOLOGICAL_PROCESS
GO:0000278; MITOTIC CELL CYCLE; BIOLOGICAL_PROCESS
GO:0000910; CYTOKINESIS; BIOLOGICAL_PROCESS
GO:0001700; EMBRYONIC DEVELOPMENT VIA THE SYNCYTIAL BLASTODERM; BIOLOGICAL_PROCESS
GO:0001745; COMPOUND EYE MORPHOGENESIS; BIOLOGICAL_PROCESS
GO:0003674; MOLECULAR_FUNCTION; MOLECULAR_FUNCTION
GO:0003676; NUCLEIC ACID BINDING; MOLECULAR_FUNCTION
GO:0003677; DNA BINDING; MOLECULAR_FUNCTION
GO:0003700; TRANSCRIPTION FACTOR ACTIVITY; MOLECULAR_FUNCTION
GO:0003702; RNA POLYMERASE II TRANSCRIPTION FACTOR ACTIVITY; MOLECULAR_FUNCTION
GO:0003704; SPECIFIC RNA POLYMERASE II TRANSCRIPTION FACTOR ACTIVITY; MOLECULAR_FUNCTION
GO:0003743; TRANSLATION INITIATION FACTOR ACTIVITY; MOLECULAR_FUNCTION
GO:0003774; MOTOR ACTIVITY; MOLECULAR_FUNCTION
GO:0003777; MICROTUBULE MOTOR ACTIVITY; MOLECULAR_FUNCTION
GO:0003779; ACTIN BINDING; MOLECULAR_FUNCTION
GO:0003924; GTPASE ACTIVITY; MOLECULAR_FUNCTION
GO:0004252; SERINE-TYPE ENDOPEPTIDASE ACTIVITY; MOLECULAR_FUNCTION
GO:0004672; PROTEIN KINASE ACTIVITY; MOLECULAR_FUNCTION
GO:0004674; PROTEIN SERINE/THREONINE KINASE ACTIVITY; MOLECULAR_FUNCTION
GO:0004702; RECEPTOR SIGNALING PROTEIN SERINE/THREONINE KINASE ACTIVITY; MOLECULAR_FUNCTION
GO:0004713; PROTEIN TYROSINE KINASE ACTIVITY; MOLECULAR_FUNCTION
GO:0004722; PROTEIN SERINE/THREONINE PHOSPHATASE ACTIVITY; MOLECULAR_FUNCTION
GO:0004725; PROTEIN TYROSINE PHOSPHATASE ACTIVITY; MOLECULAR_FUNCTION
GO:0004871; SIGNAL TRANSDUCER ACTIVITY; MOLECULAR_FUNCTION
GO:0005085; GUANYL-NUCLEOTIDE EXCHANGE FACTOR ACTIVITY; MOLECULAR_FUNCTION
GO:0005089; RHO GUANYL-NUCLEOTIDE EXCHANGE FACTOR ACTIVITY; MOLECULAR_FUNCTION
GO:0005096; GTPASE ACTIVATOR ACTIVITY; MOLECULAR_FUNCTION

GO:0005198; STRUCTURAL MOLECULE ACTIVITY; MOLECULAR_FUNCTION
GO:0005200; STRUCTURAL CONSTITUENT OF CYTOSKELETON; MOLECULAR_FUNCTION
GO:0005488; BINDING; MOLECULAR_FUNCTION
GO:0005509; CALCIUM ION BINDING; MOLECULAR_FUNCTION
GO:0005515; PROTEIN BINDING; MOLECULAR_FUNCTION
GO:0005516; CALMODULIN BINDING; MOLECULAR_FUNCTION
GO:0005524; ATP BINDING; MOLECULAR_FUNCTION
GO:0005525; GTP BINDING; MOLECULAR_FUNCTION
GO:0005575; CELLULAR_COMPONENT; CELLULAR_COMPONENT
GO:0005576; EXTRACELLULAR REGION; CELLULAR_COMPONENT
GO:0005622; INTRACELLULAR; CELLULAR_COMPONENT
GO:0005634; NUCLEUS; CELLULAR_COMPONENT
GO:0005737; CYTOPLASM; CELLULAR_COMPONENT
GO:0005739; MITOCHONDRION; CELLULAR_COMPONENT
GO:0005811; LIPID PARTICLE; CELLULAR_COMPONENT
GO:0005829; CYTOSOL; CELLULAR_COMPONENT
GO:0005834; HETEROTRIMERIC G-PROTEIN COMPLEX; CELLULAR_COMPONENT
GO:0005856; CYTOSKELETON; CELLULAR_COMPONENT
GO:0005871; KINESIN COMPLEX; CELLULAR_COMPONENT
GO:0005874; MICROTUBULE; CELLULAR_COMPONENT
GO:0005875; MICROTUBULE ASSOCIATED COMPLEX; CELLULAR_COMPONENT
GO:0005884; ACTIN FILAMENT; CELLULAR_COMPONENT
GO:0005886; PLASMA MEMBRANE; CELLULAR_COMPONENT
GO:0005938; CELL CORTEX; CELLULAR_COMPONENT
GO:0006350; TRANSCRIPTION; BIOLOGICAL_PROCESS
GO:0006355; REGULATION OF TRANSCRIPTION, DNA-DEPENDENT; BIOLOGICAL_PROCESS
GO:0006357; REGULATION OF TRANSCRIPTION FROM RNA POLYMERASE II PROMOTER; BIOLOGICAL_PROCESS
GO:0006412; TRANSLATION; BIOLOGICAL_PROCESS
GO:0006413; TRANSLATIONAL INITIATION; BIOLOGICAL_PROCESS
GO:0006468; PROTEIN AMINO ACID PHOSPHORYLATION; BIOLOGICAL_PROCESS
GO:0006470; PROTEIN AMINO ACID DEPHOSPHORYLATION; BIOLOGICAL_PROCESS
GO:0006508; PROTEOLYSIS; BIOLOGICAL_PROCESS
GO:0006810; TRANSPORT; BIOLOGICAL_PROCESS
GO:0006911; PHAGOCYTOSIS, ENGULFMENT; BIOLOGICAL_PROCESS
GO:0006915; APOPTOSIS; BIOLOGICAL_PROCESS
GO:0006952; DEFENSE RESPONSE; BIOLOGICAL_PROCESS
GO:0006955; IMMUNE RESPONSE; BIOLOGICAL_PROCESS
GO:0007010; CYTOSKELETON ORGANIZATION; BIOLOGICAL_PROCESS

GO:0007015; ACTIN FILAMENT ORGANIZATION; BIOLOGICAL_PROCESS
GO:0007017; MICROTUBULE-BASED PROCESS; BIOLOGICAL_PROCESS
GO:0007018; MICROTUBULE-BASED MOVEMENT; BIOLOGICAL_PROCESS
GO:0007049; CELL CYCLE; BIOLOGICAL_PROCESS
GO:0007052; MITOTIC SPINDLE ORGANIZATION; BIOLOGICAL_PROCESS
GO:0007067; MITOSIS; BIOLOGICAL_PROCESS
GO:0007155; CELL ADHESION; BIOLOGICAL_PROCESS
GO:0007165; SIGNAL TRANSDUCTION; BIOLOGICAL_PROCESS
GO:0007186; G-PROTEIN COUPLED RECEPTOR PROTEIN SIGNALING PATHWAY; BIOLOGICAL_PROCESS
GO:0007242; INTRACELLULAR SIGNALING CASCADE; BIOLOGICAL_PROCESS
GO:0007254; JNK CASCADE; BIOLOGICAL_PROCESS
GO:0007264; SMALL GTPASE MEDIATED SIGNAL TRANSDUCTION; BIOLOGICAL_PROCESS
GO:0007269; NEUROTRANSMITTER SECRETION; BIOLOGICAL_PROCESS
GO:0007283; SPERMATOGENESIS; BIOLOGICAL_PROCESS
GO:0007298; BORDER FOLLICLE CELL MIGRATION; BIOLOGICAL_PROCESS
GO:0007349; CELLULARIZATION; BIOLOGICAL_PROCESS
GO:0007369; GASTRULATION; BIOLOGICAL_PROCESS
GO:0007391; DORSAL CLOSURE; BIOLOGICAL_PROCESS
GO:0007399; NERVOUS SYSTEM DEVELOPMENT; BIOLOGICAL_PROCESS
GO:0007409; AXONOGENESIS; BIOLOGICAL_PROCESS
GO:0007411; AXON GUIDANCE; BIOLOGICAL_PROCESS
GO:0007417; CENTRAL NERVOUS SYSTEM DEVELOPMENT; BIOLOGICAL_PROCESS
GO:0007419; VENTRAL CORD DEVELOPMENT; BIOLOGICAL_PROCESS
GO:0007420; BRAIN DEVELOPMENT; BIOLOGICAL_PROCESS
GO:0007422; PERIPHERAL NERVOUS SYSTEM DEVELOPMENT; BIOLOGICAL_PROCESS
GO:0007423; SENSORY ORGAN DEVELOPMENT; BIOLOGICAL_PROCESS
GO:0007424; OPEN TRACHEAL SYSTEM DEVELOPMENT; BIOLOGICAL_PROCESS
GO:0007435; SALIVARY GLAND MORPHOGENESIS; BIOLOGICAL_PROCESS
GO:0007455; EYE-ANTENNAL DISC MORPHOGENESIS; BIOLOGICAL_PROCESS
GO:0007476; IMAGINAL DISC-DERIVED WING MORPHOGENESIS; BIOLOGICAL_PROCESS
GO:0007498; MESODERM DEVELOPMENT; BIOLOGICAL_PROCESS
GO:0007507; HEART DEVELOPMENT; BIOLOGICAL_PROCESS
GO:0007517; MUSCLE ORGAN DEVELOPMENT; BIOLOGICAL_PROCESS
GO:0008017; MICROTUBULE BINDING; MOLECULAR_FUNCTION
GO:0008063; TOLL SIGNALING PATHWAY; BIOLOGICAL_PROCESS
GO:0008092; CYTOSKELETAL PROTEIN BINDING; MOLECULAR_FUNCTION
GO:0008134; TRANSCRIPTION FACTOR BINDING; MOLECULAR_FUNCTION
GO:0008138; PROTEIN TYROSINE/SERINE/THREONINE PHOSPHATASE ACTIVITY;

MOLECULAR_FUNCTION
GO:0008150; BIOLOGICAL_PROCESS; BIOLOGICAL_PROCESS
GO:0008258; HEAD INVOLUTION; BIOLOGICAL_PROCESS
GO:0008270; ZINC ION BINDING; MOLECULAR_FUNCTION
GO:0008283; CELL PROLIFERATION; BIOLOGICAL_PROCESS
GO:0008293; TORSO SIGNALING PATHWAY; BIOLOGICAL_PROCESS
GO:0008340; DETERMINATION OF ADULT LIFE SPAN; BIOLOGICAL_PROCESS
GO:0008360; REGULATION OF CELL SHAPE; BIOLOGICAL_PROCESS
GO:0008407; BRISTLE MORPHOGENESIS; BIOLOGICAL_PROCESS
GO:0015031; PROTEIN TRANSPORT; BIOLOGICAL_PROCESS
GO:0016020; MEMBRANE; CELLULAR_COMPONENT
GO:0016021; INTEGRAL TO MEMBRANE; CELLULAR_COMPONENT
GO:0016055; WNT RECEPTOR SIGNALING PATHWAY; BIOLOGICAL_PROCESS
GO:0016310; PHOSPHORYLATION; BIOLOGICAL_PROCESS
GO:0016311; DEPHOSPHORYLATION; BIOLOGICAL_PROCESS
GO:0016319; MUSHROOM BODY DEVELOPMENT; BIOLOGICAL_PROCESS
GO:0016563; TRANSCRIPTION ACTIVATOR ACTIVITY; MOLECULAR_FUNCTION
GO:0016566; SPECIFIC TRANSCRIPTIONAL REPRESSOR ACTIVITY; MOLECULAR_FUNCTION
GO:0019992; DIACYLGLYCEROL BINDING; MOLECULAR_FUNCTION
GO:0030036; ACTIN CYTOSKELETON ORGANIZATION; BIOLOGICAL_PROCESS
GO:0030286; DYNEIN COMPLEX; CELLULAR_COMPONENT
GO:0030707; OVARIAN FOLLICLE CELL DEVELOPMENT; BIOLOGICAL_PROCESS
GO:0035023; REGULATION OF RHO PROTEIN SIGNAL TRANSDUCTION; BIOLOGICAL_PROCESS
GO:0035071; SALIVARY GLAND CELL AUTOPHAGIC CELL DEATH; BIOLOGICAL_PROCESS
GO:0042623; ATPASE ACTIVITY, COUPLED; MOLECULAR_FUNCTION
GO:0042803; PROTEIN HOMODIMERIZATION ACTIVITY; MOLECULAR_FUNCTION
GO:0043565; SEQUENCE-SPECIFIC DNA BINDING; MOLECULAR_FUNCTION
GO:0045449; REGULATION OF TRANSCRIPTION; BIOLOGICAL_PROCESS
GO:0048102; AUTOPHAGIC CELL DEATH; BIOLOGICAL_PROCESS
GO:0048477; OOGENESIS; BIOLOGICAL_PROCESS
GO:0048666; NEURON DEVELOPMENT; BIOLOGICAL_PROCESS
GO:0048749; COMPOUND EYE DEVELOPMENT; BIOLOGICAL_PROCESS
GO:0048813; DENDRITE MORPHOGENESIS; BIOLOGICAL_PROCESS
GO:0051726; REGULATION OF CELL CYCLE; BIOLOGICAL_PROCESS

Supplementary Table 4. GO categories for GSEA. A minimum of fifteen genes was required for inclusion as a gene set for GSEA. The 136 GO annotations containing at least fifteen

Drosophila genes are listed here. See text for additional discussion of input data and parameters for GSEA.

Chapter 5:
Conclusion

In this thesis, we have developed and validated methods for the analysis of high-throughput single-cell morphological data from genetic screens. We applied our methods to the genes and morphological processes involved in cell locomotion, for the reason that gaining an increased understanding of cell locomotion is essential to understanding metastatic disease. Indeed, in order to study signaling in locomotion and metastasis, we developed statistical techniques to identify genetic components of cellular morphological variability, to utilize morphological data to indentify genetic interactions and perform signaling pathway inference, and to integrate morphological and transcriptional data to study genetic contributions to cell morphology.

Genetic Contributions to Variability (Chapter 2)

Summary

Our goal in this chapter was to define and apply robust statistical measures to identify genes regulating morphological variability. We introduced methods for measuring genetic contributions to morphological variability for specific cellular processes. We developed a robust method for measuring population variability more generally, and applied this method to genetic screens in both yeast and fly. The basis for the metric was relatively simple: a multi-dimensional

analog of one-dimensional variance, applied to data that has been normalized by taking z-scores in each raw dimension and then reduced in dimensionality by using PCA. The benefit of applying a relatively simple methodology was increased confidence in the interpretation of results.

We applied our variability scoring procedure to study genetic contributions to morphological variability in specific cellular processes (protrusion/adhesion formation and adhesion disassembly/cortical tension in fly; septin ring formation in yeast). We validated our results based on known gene functions and network architectures for the processes under consideration. We found that the effects of genetic perturbations on morphological variability were explicable in many situations by the network architecture of the cellular process under consideration. Our results extended the finding of Levy et al. [1] that knockout of network hubs tends to increase morphological variability. Here, we considered more intricate network architectures associated with regulation of complex cellular processes. Indeed, work in measuring single-gene transcription has shown that perturbation of expression of genes with upstream products causes increased noise in the expression of downstream targets [2,3]. We demonstrated repeatedly that perturbation of genes acting upstream in signaling pathways tended to increase morphological noise in the process mediated by the pathway to a greater extent than perturbation of genes acting further downstream in the pathway.

We proposed that genes resulting in populations with high morphological variability when inhibited by RNAi are *suppressors of noise* , and conversely, that genes driving populations towards abnormal homogeneity when inhibited are *enhancers of noise*. To study suppressors and enhancers of morphological noise, we utilized TCs from the *Drosophila* screen thought to be involved in regulating a single morphological process, called phenoclusters [4], for example

protrusion/adhesion formation. By comparing variability p-scores across TCs within a functionally-related cluster, we identified genes that modulate morphological noise for a particular morphological process. We identified suppressors and enhancers of noise for protrusion/adhesion formation and adhesion disassembly/cortical tension in fly that were consistent with the known architecture of RhoGTPase signaling. Likewise, in yeast, we studied the process of septin ring formation, identifying new roles for genes as suppressors or enhancers of morphological noise, and showing that variability of knockout TCs was consistent with the regulatory architecture of septin ring formation.

In addition to studying genetic contributions to morphological variability within particular cellular processes, this chapter proffered an alternative method for measuring variability to that of Levy et al. Their method was effective at identifying genes which, when knocked out, resulted in highly variable morphology in a great many feature dimensions in yeast. We showed that, for yeast, our metric yielded similar results to theirs in the case of TCs with highly variable morphology. But above and beyond that, our method was effective at studying more subtle changes in morphological variability in cellular processes. It was also effective at identifying genes which, when knocked out, result in decreased variability. And we applied it to *Drosophila* cell culture as well as yeast.

In sum, we developed methods to probe the genetic basis of morphological stochasticity. Because ours were the first methods to quantitatively study the genetic regulation of morphological variability in cellular processes, we validated our results using known functional properties of genes in our datasets and regulatory architecture of relevant cellular processes. As additional sources of morphological data (different cell lines, different organisms) become available, we expect these methods to serve to quantify morphological variability of single-cell

populations and genetic modulation of morphological stability and variation in cellular processes.

Future Work

Further work on morphological variability will study the effects of different regulatory structures on modulation of variability by detailed study of other cellular processes. At present, this is limited by the fact that certain cellular processes tend to produce more much dramatic morphological changes than others. For example, our methods do not find significant results for lamellipodia formation in fly. Further refinement of our methods, as well as new genetic screens to obtain image-based data, will be necessary to increase sensitivity and study less dramatically modulated cellular processes in a meaningful way.

Additional work will attempt to incorporate the informative signal from morphological variability into models for network inference. Indeed, we demonstrated that knockouts of upstream proteins tends to increase observed population variability p-scores while maintaining relatively similar mean morphology as compared to knockouts of downstream targets. Such a framework would need to be combined with other data sources (e.g. transcriptional, proteomic) to obtain meaningful results. But the key to using morphological data to obtain directionality in signaling networks without using any *a priori* knowledge of the underlying network structure is to use measurement of population-level variability, as this capitalizes on the noise propagation properties of signaling pathways, as demonstrated repeatedly in this paper. On the other hand, if one uses *a priori* knowledge of this sort, then it is possible to directly infer signaling relationships by comparing similarity of the morphology of upstream and downstream targets–

although it requires double-knockout data, as described thoroughly in Chapter 3 and which we now review.

Signaling Pathway Inference (Chapter 3)

Summary

In this chapter our aim was to perform inference of protein signaling relationships by utilizing high-throughput morphological data. Toward this goal, we first developed a systematic framework for identifying genetic interactions on the basis of high-throughput (single- and double-knockout) morphological data from an RNAi screen. We then applied this framework to infer RhoGAP/GTPase regulatory relationships by using prior knowledge of the basic structure of RhoGAP/GTPase signaling.

We defined a classification model for assigning a set of putative upstream TCs to a set of putative downstream TCs using a voting scheme. This model was used to classify GAP single-knockout TCs onto the set of GTPase overexpression TCs. This analysis was repeated for double-knockout experiments. The classification model allowed us to assign any new point, or set of points, in morphological space to one of several classes. First, we used GTPase overexpression experiments as the downstream classes, and for each GAP single-knockout, we used the model to classify that knockout as belonging to one of downstream classes. Second, we used GTPase overexpression experiments and GAP double-knockouts in an analogous way. This allowed us to associate to each GAP a GTPase which it is most likely to regulate. These predictions were compared to biologically validated interactions and non-interactions between

GAPs and GTPases. We also tested various other methods for performing inference using double-knockout data, and showed that our primary classification model outperforms these alternatives. In addition, we developed a similar classification model using single knockouts as the set of "downstream" targets and double-knockouts as the set of "upstream" targets. By systematically identifying double-knockouts TCs that are morphologically similar to single-knockout TCs, we were able to identify genetic interactions between the GAPs and to construct putative hierarchies of action for GAPs.

This section of the thesis made several fundamental contributions to the field. The first contribution was to develop methods for using high-throughput morphological data in a systematic fashion to identify genetic interactions. Second, we showed the fundamental fact that with additional prior knowledge of the basic network structure, our framework can be used to identify signaling interactions successfully. Third, the computational framework presented here represented an initial approach to the problem that will serve as a basis for future enhancements (see below). Fourth, and perhaps most intriguing, we showed that our classification model performs much better with both single- and double-knockout data versus only single-knockout data.

On the latter point, each GAP likely regulates multiple GTPases and each GTPase is likely regulated by multiple GAPs. This means that a knockout of a single GAP may not robustly increase activity of any one of the GTPases it normally regulates. However, knockout of two GAPs, each normally regulating the same GTPase, more likely results in increase activity of that GTPase. Put simply, because the regulatory structure is redundant, combinatorial knockouts are necessary to generate a sufficiently informative signal for successful prediction. Our finding in

the context of morphological data parallels that of phosphoproteomics data, for which the power of utilizing double-knockouts has been demonstrated [5].

Future Work

Future work will involve acquisition of additional double-knockout morphological data to allow for prediction of other known GTPase targets, as well as for better simultaneous predictions of multiple GTPase targets for a single GAP knockout. For the latter task, one possibility would be to obtain double-overexpression GTPase data and augment the classification model with these TCs as targets. A GAP treatment condition mapped to a double-overexpression class (versus either of the single overexpression classes) would suggest multiple GTPase targets for the GAP. Additional work will involve application of our methods to new image-based data sources, including image-based data related to pathways that are less redundant, for example VEGF (PVR) and MAPK pathways [6, 7].

Integration with Transcriptional Data (Chapter 4)

Summary

In this chapter, we aimed to integrate expression data with high-throughput morphological data to study the mechanisms for determination of cell morphology. We utilized the morphological data from the 273-TC *Drosophila* genetic screen as well as microarray data from a similar screen. By comparing expression data between control treatment conditions and treatment

conditions displaying a particular morphological phenotype of interest (e.g. high population variability), we identified genes and pathways correlated with this class distinction, thereby validating our previous studies and providing a means for generating new genes of interest for future study.

Our framework required the definition of one or more class distinctions to separate treatment conditions into pairs of classes on the basis of morphology and, subsequently, to use transcriptional data to determine differential expression between these pairs. We generated two different types of class distinctions corresponding to phenoclusters and variability analysis (in the last case, building off the results of Chapter 2). More specifically, we considered class distinctions defined by: phenoclusters versus control; and high/low morphological variability versus control and high versus low morphological variability.

For each class distinction, we selected all treatment conditions from the *Drosophila* BG-2 morphology screen also present in a *Drosophila* microarray screen in the S2R+ cell line [8] that fall into the two groups dictated by the class distinction; we then determined genes that were differentially expressed as well as pathways that were enriched between the two groups. After normalizing the microarray data, we performed t-tests as well as significance analysis of microarrays (SAM) to determine differential expression, and we carried out gene set enrichment analysis (GSEA) to determine gene set enrichment.

Differential expression of single genes was essentially absent when using standard methods based on t-tests and correction for multiple hypothesis testing. Using a less stringent method (SAM), it was possible to identify single genes exhibiting moderate differential expression for some of the class distinctions under consideration. In many cases, individual genes that were

identified by this analysis can be rationalized with the relevant class distinction, for example up-regulation of the RhoGEF CG30440 in the case of high morphological variability. In the course of our analysis, several genes of unknown function were determined to be differentially expressed (by SAM) across the morphological class distinctions under consideration, thus generating potential targets for future biological experimentation. Gene set enrichment analysis produced substantive results that provide strong evidence that our methods are successfully detecting informative signals. For example, for the class distinction defined by high versus low morphological variability, expression levels for the mTOR pathway, which is strongly associated with translational control, stress response, and locomotion, were enriched for the high-variability treatment conditions. For the distinction defined by control versus the Lamellipodia Formation Phenocluster, the Wnt pathway, VEGF pathway, cell cycle category, and gastrulation category were all found to be down-regulated in the TCs showing morphology similar to cells unable to form lamellipodia. These findings were consistent with known biology – for example, the Wnt pathway is implicated in lamellipodia formation. Numerous other examples of biologically validated findings were described in Chapter 4, and provide strong motivation for further pursuing this line of research.

Future Work

As previously noted, different cell lines were used for morphological and transcriptional data (BG-2 for morphology, S2R+ for transcription). This may help explain the lack of significant results when using t-tests, as signal strength is diminished when comparing alternate cell lines. On the other hand, because we did obtain meaningful results even when comparing different cell

lines, we are encouraged to carry out further experiments to continue this line of research in future work – namely, to obtain microarray data for a screen using BG-2 cells. The framework we used here can be applied to the BG-2 microarray and BG-2 morphology data in order to study the genetic of morphological processes with far greater precision than when using different cell lines. In order to further study morphological variability as a cellular phenotype, the acquisition and integrative analysis of BG-2 microarrays for treatment conditions yielding extreme high/low variability scores should be carried out. The combination of transcriptional and morphological data will generate new insights into the determination of morphology.

References

1. Levy SF, Siegal ML. Network Hubs Buffer Environmental Variation in Saccharomyces cerevisiae. PLoS Biol. 2008 Nov 4;6(11):e264.

2. Pedraza JM, van Oudenaarden A. Noise propagation in gene networks. Science. 2005 Mar 25;307(5717):1965-9.

3. Raj A, van Oudenaarden A. Nature, nurture, or chance: stochastic gene expression and its consequences. Cell. 2008 Oct 17;135(2):216-226.

4. Bakal C., Aach J., Church G., Perrimon N.: Quantitative Morphological Signatures Define Local Signaling Networks Regulating Cell Morphology. Science 316(5832), 1753-1756 (2007)

5. Bakal C, Linding R, Llense F, Heffern E, Martin-Blanco E, Pawson T, Perrimon N. Phosphorylation Networks Regulating JNK Activity in Diverse Genetic Backgrounds. Science. 2008 Oct 17;322(5900):453-456.

6. Kiger A, Baum B, Jones S, Jones M, Coulson A, Echeverri C, Perrimon N. A functional genomic analysis of cell morphology using RNA interference. Journal of Biology. 2003 ;2(4):27.

7. Sims D, Duchek P, Baum B. PDGF/VEGF signaling controls cell size in Drosophila. Genome Biol. 2009 Feb 12;10(2):R20.

8. Baym M, Bakal C, Perrimon N, Berger B.: High-Resolution Modeling of Cellular Signaling Networks. Proceedings of the 12th Annual International Conference on Research in Computational Molecular Biology (RECOMB 2008), LNBI 4955: 257-271, 2008